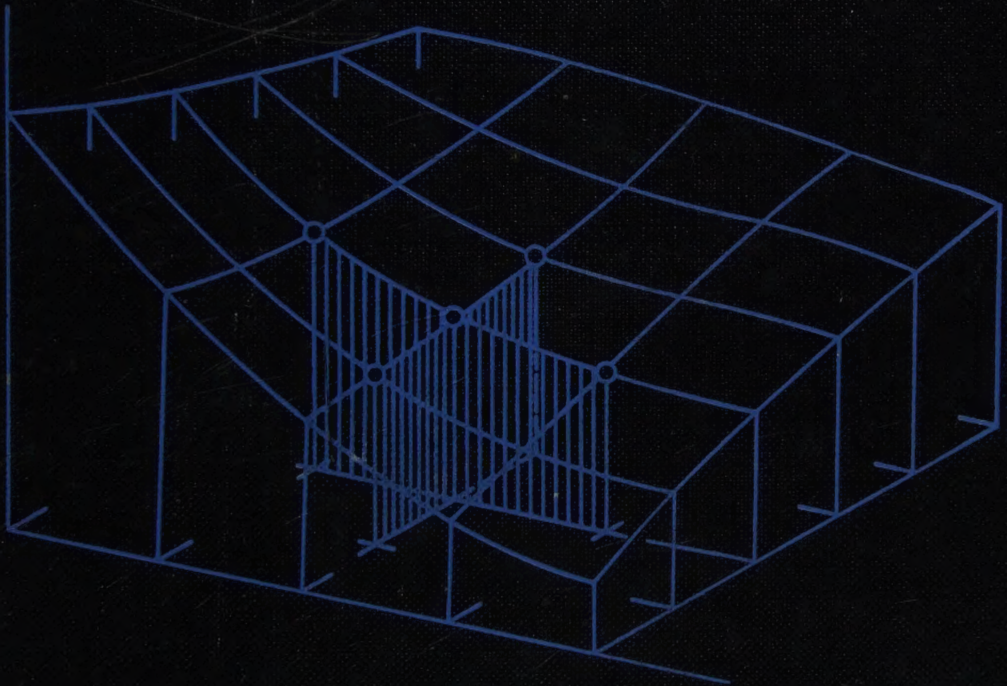
modern methods of engineering computation



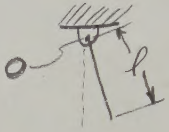
Robert L. Ketter/Sherwood P. Prawel, Jr.

MASS MOMENTS of INERTIA ~ J

① $J = \frac{1}{2} m R^2$ HOMOGENEOUS

② $J = 2/5\, m R^2$ SPHERE

③ $J_o = \frac{1}{3} m \ell^2$



SLENDER ROD
ABOUT END

$$J_o = J_{cg} + mL^2$$

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

$2/50

modern methods of engineering computation

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

modern methods of engineering computation

Robert L. Ketter, Professor
Faculty of Engineering and Applied Science
State University of New York at Buffalo

Sherwood P. Prawel, Jr., Associate Professor
Faculty of Engineering and Applied Science
State University of New York at Buffalo

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

McGraw-Hill Book Company
New York St. Louis San Francisco London
Sydney Toronto Mexico Panama

Modern Methods of Engineering Computation

Copyright © 1969 by McGraw-Hill, Inc. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

Library of Congress Catalog Card Number 72-81608
ISBN 07-034423-x

9 10 11 12 13 14 15 16 KPKP 78321098

preface

It is to be recognized and understood at the outset that this work is intended to be an *introduction* to the field of modern computational methods. It is not designed to be, nor does it constitute, a definitive treatment of any one or more of the various topics that make up this extremely broad area of study. Equally true, however, it is not a superficial, topical survey of the subject. It is a serious attempt to develop from first principles the various computational methods that are basic and/or are in general use today, and to indicate the interrelationships that exist among them. It also is hoped that the treatment is meaningful in terms of problem solving, since this was one of our guiding purposes throughout its preparation.

The major parts of the text have been written with second- or third-year students of engineering and applied science in mind—students who need a general "overview" of numerical procedures to facilitate upper-level course work. It must be noted, however, that certain beginning graduate students also have found it meaningful and have obtained approval for its inclusion in their master's degree programs.

During the past five to six years, over twelve hundred students from all branches of engineering, from mathematics, and from physics, at all levels from sophomores through first-year graduate students, have been "subjected" to this particular presentation. Their enthusiastic response to the general approach used has in large measure kept us plodding toward eventual publication, even though at times we have been tempted to give it last order of priority. Their comments and suggestions for improvement and clarification have been most helpful, and are sincerely appreciated.

Possibly the most troublesome problem with which we were faced was our desire to produce a text which could, by the inclusion or exclusion of certain chapters and/or sections of chapters, be taken either before or after a more "department of mathematics" oriented course in numerical analysis. We admit that we have only partially accomplished this. Student reaction in this regard, however, has been more favorable. The course has generated an interest in the general area of numerical computations, and many of the students who have elected to take further courses in this area have indicated positively their feelings toward this first presentation. Equally true, upper-level students in mathematics who have taken this course following a more rigorous mathematical treatment have frequently exclaimed that for the first time they "see" how it fits together, and how the methods can be used to solve real problem situations.

It is impossible in a text of this kind to present a level of treatment that is sufficiently intensive or rigorous enough to satisfy the needs or desires of all students who have a serious interest in any one of the subject areas. As noted, the primary purpose of this text is to provide an "overview." Therefore, for supplementary but not necessarily secondary reading, there is included a selected bibliography of reference texts.

Finally, the authors wish to express their appreciation to all who have directly assisted them in making this work possible: to Charles Fogel, Vipperla Venkayya, Jack Blackmon, Andrew Frankus, and others who have made many valuable suggestions concerning the order and clarity of presentation; to Dwane Moore, Edward Colman, Robie Button, and others who checked and rechecked the problems and their solutions; and to Mary Basehart, Edna Thill, Lucy Bianchi, and Clara Baideme for typing the text. To all we acknowledge our gratitude.

ROBERT L. KETTER
SHERWOOD P. PRAWEL, JR.

contents

Preface v

Selected References xiii

Chapter ONE
Introduction **1**

1.1 General Engineering Problems 1
1.2 The Engineering Approach 2
1.3 Errors 3
1.4 Presentation of Engineering Problems 3

Chapter TWO
Digital Computers **5**

2.1 Introduction: Computers 5
  2.1.1 Digital Computers 6
  2.1.2 Analog Computers 6
2.2 The Computer Program 7
2.3 The Flow Diagram 9

Chapter THREE
Determinants **24**

3.1 Introduction 24
3.2 Evaluation of a Determinant 25
3.3 Special Properties of Determinants 27

3.4 Row or Column Addition 28
3.5 Pivotal Condensation 29
3.6 Flow Diagram for Pivotal Condensation 31

Chapter FOUR
Matrices **40**

4.1 Introduction 40
4.2 Special Matrices 43
    4.2.1 The Square Matrix 43
    4.2.2 The Diagonal Matrix 43
    4.2.3 The Unit or Identity Matrix 44
    4.2.4 The Upper Triangular Matrix 44
    4.2.5 The Upper Unit Triangular Matrix 44
    4.2.6 The Lower Triangular Matrix 44
    4.2.7 The Null or Zero Matrix 44
    4.2.8 The Transpose of a Matrix 45
4.3 Elementary Matrix Operations 45
    4.3.1 Matrix Addition 45
    4.3.2 Matrix Multiplication 46
4.4 Special Considerations Concerning Matrix Multiplication 47
    4.4.1 Multiplication by the Identity Matrix 48
    4.4.2 Matrix Multiplication Is Associative 48
    4.4.3 Matrix Multiplication Is Not in General Commutative 48
    4.4.4 Matrix Multiplication Is Distributive 49
    4.4.5 Multiplication by a Constant 49
4.5 Partitioning of Matrices 49
4.6 Flow Diagrams for Elementary Matrix Operations 62
    4.6.1 Matrix Addition 63
    4.6.2 Multiplication of a Matrix by a Constant 64
    4.6.3 Matrix Multiplication 64

Chapter FIVE
Simultaneous Linear Algebraic Equations **66**

5.1 Introduction 66
5.2 Cramer's Rule 69
5.3 Elimination Methods 71
    5.3.1 Gauss' Method 71
    5.3.2 Cholesky's Method 76
    5.3.3 Square-root Method 81
5.4 Iteration Methods 83
    5.4.1 Jacobi's Method 84
    5.4.2 Gauss-Seidel Method 85
    5.4.3 Iteration of Nondiagonal Sets 87
5.5 Relaxation Methods 87
    5.5.1 Relaxation of 3 × 3 Set of Equations 89
    5.5.2 Relaxation of Larger Sets of Equations 93
5.6 Computer Solutions 97
    5.6.1 Gauss-Seidel Iteration 98
    5.6.2 Cholesky's Method 102

Chapter SIX
The Inverse Problem **118**

6.1 Introduction 118
6.2 Cramer's Rule 119

6.3 Elementary Row Operators 122
6.4 Elimination Methods 123
  6.4.1 Gauss Elimination 126
  6.4.2 Cholesky's Method 127
6.5 Inversion of Matrices by Iteration or Relaxation 128
6.6 Matrix Inversion by Partitioning 128
6.7 Computer Solutions 133

Chapter SEVEN
Eigenvalue Problems 136

7.1 Introduction 136
7.2 Examples of Eigenvalue Problems 139
7.3 Determination of Eigenvalues by Iteration 145
7.4 Modification of the Coefficient Matrix [*B*] 147
7.5 Computer Solutions 150

Chapter EIGHT
Nonlinear Equations 158

8.1 Introduction 158
8.2 Methods for Obtaining First Approximations 159
  8.2.1 Graphical Methods 159
  8.2.2 Special Methods for Approximating the Real Roots of Nonlinear Algebraic Equations 162
  8.2.3 Relationships that Exist between Roots of Algebraic Equations 164
  8.2.4 Synthetic Division 166
8.3 Methods for Improving the Accuracy of a Trial Solution 168
  8.3.1 False-position Method 168
  8.3.2 Newton's Method of Tangents 170
  8.3.3 Newton's Second-order Method 171
  8.3.4 Direct Iteration 174
8.4 Real Roots of Transcendental Equations 175
8.5 Complex Roots of Nonlinear Algebraic Equations 177
8.6 Real Roots of Simultaneous Nonlinear Algebraic Equations 180
  8.6.1 Newton's Method 181
  8.6.2 Direct Iteration 182
8.7 Computer Solutions 186

Chapter NINE
Miscellaneous Methods 188

9.1 Difference Tables 188
9.2 Errors in Difference Tables 195
9.3 The Difference Operators Δ, ∇, and δ 197
9.4 Interpolation 204
9.5 Inverse Interpolation 218
9.6 Numerical Differentiation and Integration 220
  9.6.1 Difference Expressions 222
  9.6.2 Integration 229
9.7 Least-squares Approximations 233
9.8 Computer Solutions 241
  9.8.1 Construction of a Forward-difference Table 241
  9.8.2 Development of the Lagrange Interpolation Formula 243
  9.8.3 Numerical Integration Using Simpson's One-third Rule 244

SKIPPED

Chapter TEN
Numerical Solution of Ordinary Differential Equations 245

10.1 Introduction 245
10.2 Taylor Series Expansion: Initial-value Problems 247
    10.2.1 First-order Differential Equations 247
    10.2.2 Sets of First-order Differential Equations 252
10.3 Series Determination by Iteration of Integral Equations: Initial-value Problems 253
    10.3.1 First-order Differential Equations 253
    10.3.2 Sets of First-order Differential Equations 254
10.4 Polynomial Approximations 255
    10.4.1 Collocation 256
    10.4.2 Subdivision Method 257
    10.4.3 Least-squares Method 257
    10.4.4 Galerkin's Method 258
10.5 Higher-order Differential Equations: Initial-value Problems 259
10.6 Euler Methods of Forward Integration 262
    10.6.1 Euler's Method 262
    10.6.2 Modified Euler-Cauchy Method 263
10.7 Runge-Kutta Methods 268
    10.7.1 First-order Equations 268
    10.7.2 Sets of First-order Equations 275
    10.7.3 Higher-order Equations 278
10.8 Analytical Continuation 280
    10.8.1 Predictor-Corrector Methods 281
    10.8.2 Collocation Method 285
10.9 Stability of Numerical Solutions to Initial-value Problems 286
10.10 Boundary-value Problems 287
10.11 Eigenvalue Problems 296
10.12 Nonlinear Boundary-value Problems 301

Chapter ELEVEN
Numerical Solution of Partial Differential Equations 318

11.1 Introduction 318
11.2 Two-dimensional Grids 320
11.3 Development of Finite-difference Computational Molecules 323
    11.3.1 "Skewed"-coordinate Computational Molecules 326
    11.3.2 Triangular-coordinate Molecules 333
    11.3.3 Polar Coordinates 335
11.4 Irregular Boundaries 335
11.5 Elliptic Partial Differential Equations: Boundary-value Problems 343
11.6 The Use of Relaxation Methods in the Solution of Boundary-value Problems 359
    11.6.1 Unit Relaxation 359
    11.6.2 Group Relaxation 364
    11.6.3 Change in Grid Size 371
11.7 Parabolic Partial Differential Equations: Transient Problems 379
    11.7.1 Explicit Formulation 380
    11.7.2 Implicit Formulation 395
    11.7.3 Method of Lines 400
11.8 Hyperbolic Partial Differential Equations 401
    11.8.1 Method of Characteristics 401
    11.8.2 Finite-difference Method 418

Chapter TWELVE
Optimization 427

12.1 Introduction: Concepts and Definitions 427
  12.1.1 Constrained Maximum and Minimum Problems 431
  12.1.2 Other Classifications 433
12.2 Lagrangian-multiplier Method 433
12.3 Introduction to the Calculus of Variation 438
  12.3.1 Ritz's Method 449
  12.3.2 Galerkin's Method 452
12.4 Direct vs. Indirect Methods 455
12.5 Single-variable Search 456
  12.5.1 The Search Plan 457
  12.5.2 The Golden Section 460
  12.5.3 Fibonacci Search 462
12.6 Multivariable Search 466
  12.6.1 The One-at-a-time, or Sectioning, Method 466
  12.6.2 The Method of Steepest Descent (or Ascent) 467

Index 483

selected references

Bramble, J. H. (ed.): "Numerical Solutions of Partial Differential Equations," Academic Press Inc., New York, 1966.

Collatz, L.: "Numerical Treatment of Differential Equations," 3d ed., Springer-Verlag OHG, Berlin, 1960. (Translated from a supplemental version of the second German edition by P. G. Williams.)

Dorn, W. S., and H. J. Greenberg: "Mathematics and Computing: With Fortran Programming," John Wiley & Sons, Inc., New York, 1967.

Forsythe, G. E., and W. R. Wasow: "Finite Difference Methods for Partial Differential Equations," John Wiley & Sons, Inc., New York, 1960.

Fox, L.: "Numerical Solutions of Ordinary and Partial Differential Equations," Pergamon Press, New York, 1962.

———: "An Introduction to Numerical Linear Algebra," Oxford University Press, Fair Lawn, N.J., 1965.

Greenspan, D.: "Introductory Numerical Analysis of Elliptic Boundary Value Problems," Harper & Row Publishers, Incorporated, New York, 1965.

Hamming, R. W.: "Numerical Methods for Scientists and Engineers," McGraw-Hill Book Company, New York, 1962.

Henrici, P.: "Elements of Numerical Analysis," John Wiley & Sons, Inc., New York, 1964.

Hildebrand, F. B.: "Introduction to Numerical Analysis," McGraw-Hill Book Company, New York, 1956.

Kelly, L. G.: "Handbook of Numerical Methods and Applications," Addison-Wesley Publishing Company, Inc., Reading, Mass., 1967.

Kreyszig, E.: "Advanced Engineering Mathematics," 2d ed., John Wiley & Sons, Inc., New York, 1967.

Leitmann, G. (ed.): "Optimization Techniques," Academic Press Inc., New York, 1962.

McCormick, J. M., and M. G. Salvadori: "Numerical Methods in Fortran," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

Milne, W. E.: "Numerical Solutions to Differential Equations," John Wiley & Sons, Inc., New York, 1953.

Milne-Thompson, L. M.: "Calculus of Finite Differences," St. Martin's Press, Inc., New York, 1951.

Ralston, A.: "A First Course in Numerical Analysis," McGraw-Hill Book Company, New York, 1965.

———, and H. S. Wilf (ed.): "Mathematical Methods for Digital Computers," John Wiley & Sons, Inc., New York, 1960.

Shaw, F. S.: "Relaxation Methods," Dover Publications, Inc., New York, 1953.

Todd, J. (ed.): "Survey of Numerical Analysis," McGraw-Hill Book Company, New York, 1962.

Wilde, D. J.: "Optimum Seeking Methods," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

modern methods of engineering computation

introduction

1

CLASSIFICATION

1 a) STEADY STATE
b) TRANSIENT } PROBLE

2 a) DISCRETE-ELEMENT
b) CONTINUOUS } SYS

1.1 General Engineering Problems

1) Problems in engineering can be conveniently classified as a) steady-state or b) transient in nature. Typical steady-state problems, often referred to as equilibrium problems, are equilibrium stresses in structures, steady-state temperature problems, steady-state fluid flow, and steady-state flow of direct and alternating currents. Electromagnetic energy fields, propagation of heat, and propagation of pressure waves in a fluid are all typical transient problems.

In certain cases, because of particular conditions imposed by the governing equations, an instantaneous change from one given situation to another may occur. This phenomenon, which does not conveniently fit into the steady-state-versus-transient classification given above, is referred to as an eigenvalue or characteristic-value type of problem. Buckling of structures, natural frequency of vibration, and resonance in acoustics and electric circuits are typical eigenvalue problems.

2) Engineering problems also can be categorized according to their physical form as a) discrete-element systems (those consisting of a finite number of degrees of freedom) or as b) continuous systems. The steady-state forced vibration of a given number of spring-connected masses is the discrete-element type. In terms of the amplitudes of the various

masses, solution of this problem can be reduced to solution of a set of linear simultaneous algebraic equations. Solution of the transient problem for this same mass-spring system corresponds to the solution of a set of simultaneous ordinary differential equations with prescribed initial conditions.

An example of a steady-state problem in a continuous system is the deformation of a plate with prescribed edge conditions and subjected to a nonvarying lateral load. Here, the solution is defined by partial differential equations with *closed boundary conditions* (a boundary-value problem). Transient heat flow in a three-dimensional medium requires the solution of partial differential equations with some *open boundary conditions* (an initial-and-boundary-value problem).

Certain cases that are observed to be physically continuous may actually be solved as if they were distinct discrete elements. The steady-state flow of fluid in a pipe network is continuous in physical form; however, if the rates of fluid flow in each of the various segments are considered to be the unknowns, the resulting equations are—to all intents and purposes—those associated with a distinct-element system. Similarly, it may be possible in many other cases to group unknown parameters in a manner such that the resulting "lumped systems" are the discrete-element type. Finite-difference mathematical procedures are also used to obtain approximations of continuous systems.

1.2 The Engineering Approach

While the equations considered by both the engineer and the mathematician may be of the identical type and form, the attitude and approach to a problem by the two groups may be markedly different. The solution of a differential equation by a mathematician first requires an examination of the existence of a solution. Next, properties of the solution are considered. Finally, development of a suitable form for the resulting equation(s) is undertaken. The engineer, on the other hand, when faced with a physical situation must derive first the governing differential equation(s). The actual existence of a solution to the physical problem often is taken for granted since the physical problem does exist. The existence proof of a solution to the mathematical model that has been selected to represent the physical problem is also often skipped by the engineer. Having obtained an answer, within the accuracy required, an engineer should then go back and assure himself that the solution satisfies the physical problem originally considered.

Many engineering problems do not lend themselves readily to closed-form analytic solutions. In such cases, approximations are required. These may take one of several forms: (1) series solutions; (2) direct numerical solution of the differential equations, using whatever computational facilities may be available; and (3) solution by analog methods.

This volume is intended to be an introduction to the general area of engineering computations. The presentation has been slanted purposely toward application of computational methods to the solution of problems in engineering and applied science.

1.3 Errors

It should be understood at the outset that most numerical calculations are by their very nature inexact. Primarily, the errors are due to one of three sources: inaccuracies in the original data, lack of precision in carrying out elementary operations, and inaccuracies introduced by approximate methods of solution.

Of particular significance are the errors due to "roundoff" and the inability of carrying, in a given calculation, more than a certain number of significant figures. Terms such as "absolute error," "relative error," and "truncation error" have a very real meaning, and frequently an analysis parallel to one in question must be carried out to establish the reliability of a given answer.

The problem of error analysis due to roundoff was not considered to be too difficult when calculations were carried out by hand or by desk calculators. Added places and/or error terms could be introduced with little additional work. However, there is evidence that available elementary "error theories" do not seem to be adequate for estimation, with any real degree of certainty, of the roundoff and truncation errors that result when modern high-speed digital computers are used. As discussed in the sections on digital computations, *range arithmetic* and *double or multiprecision solutions* may be required to afford an estimate of the reliability of a given answer.

1.4 Presentation of Engineering Problems

No particular form or method of presentation distinguishes an engineering problem from any other problem in applied mathematics. Single and coupled, ordinary and partial differential equations with initial boundary values, are encountered, as also are integral equations. The need to solve sets of linear and nonlinear simultaneous algebraic equations, to optimize with respect to a given condition the accumulative effects of a large number of arbitrary variables, and the need to determine the roots of transcendental and algebraic equations are also frequently encountered.

A convenient form for the presentation and discussion of a large number of engineering problems, which is now well established in up-to-date undergraduate mathematics study for engineering students, is matrix analysis. Because of the importance of this topic, Chap. 4 is devoted entirely to its discussion. Extensive use of basic matrix theory is made in certain of the methods of solution of simultaneous equations and in obtaining the inverses of linear systems. Eigenvalue problems are also approached from a matrix point of view.

digital computers

2

2.1 Introduction: Computers

Within the past ten years, profound changes have taken place in the engineer's attitude toward, and his methods of approach to, the solution of many types of problems. These changes have been due primarily to the widespread availability of various types of mechanical and electrical computing machines. These high-speed devices can be *programmed* to carry out the many routine computational tasks required for problem solution, and also to make certain simple decisions. This automation of the computational drudgery has allowed more time to be devoted to improvements in problem formulation. It is in this area that most attention is currently being focused.

Three general classes of computers are in use today: the digital computer, the analog computer, and the hybrid computer. As its name implies, the digital computer operates with *digits* or numbers. The analog computer, on the other hand, entails use of continuous variables which are related to the actual variables of the problem. The hybrid computer combines the best of both the digital and the analog computer.

2.1.1 Digital Computers

Most digital computers have the capacity to do only arithmetic. More specifically, their mode of computational operation is algebraic addition. However, by properly programming the sequence of additions, these machines can be made to perform other operations, for example, multiplication and division. Nevertheless, it should be noted that since every operation must be reduced to the basic form of addition, digital computers, by their very nature, are extremely poor "mathematicians." They are, however, very rapid in performing these additions and can carry a large number of significant figures, thereby compensating for most of their "mathematical limitations."

To solve a particular problem by use of the digital computer, instructions concerning the sequence of operations to be carried out must in some fashion be transmitted to the machine. In the case of a desk calculator, these instructions are usually contained in the operator's mind. In an electronic digital computer, however, the set of instructions must be "entered" into the machine, stored, and acted upon with little or no interruption by the operator. Since, as pointed out above, this type of computer can do only basic arithmetic, all instructions must use only basic arithmetic operations. Approximations to rigorous solutions must be used when operations other than these are required. Because of these approximations and of the other errors that were itemized in Sec. 1.3, the accuracy of digital-computer solutions may sometimes be limited.

It is apparent that normal usage of a digital computer requires of the operator a knowledge of the most exacting detail relative to the method of solution to his problem. He must also be able to transmit his thoughts to the machine.

2.1.2 Analog Computers

The analog computer differs from the digital computer in that it makes use of continuous variables rather than digits. The variables of the computer are in some way related to the variables of the problem to be solved. That is, measurements taken of variables on the analog computer are to some scale proportional to variables in the real system. The input and the output for this type of machine are also of a continuous nature, representing a real function. Very often the output from an analog computer is a graph in which the variable in question is plotted as a function of time.

The analog computer is a very fine "mathematician." It does not approximate the solution of a given problem numerically as does the digital computer. Rather, it solves the given continuous system rigorously. It should be noted, however, that accuracy on an analog computer is usually limited to no more than four significant figures. This limitation in accuracy is due to deficiencies both in the computing components and in the readout equipment. If this degree of accuracy can be tolerated—and it should be noted that in many cases this is better than the accuracy of the engineering data—the analog-computer method of solution has much to offer. In fact, in some instances, use of the analog computer is superior to use of the digital computer. For example, it is not necessary to know how to solve a particular problem to be able to obtain a solution with an analog computer. The only requirement is that the problems be correctly formulated. Programming for an analog-computer solution can therefore be much less difficult than programming the same problem for digital-computer solution.

Two general types of analog computers are in use today: those of a direct analog nature and those of an indirect or mathematical analog nature. The *direct analog computer* is usually a special-purpose device made to handle a specific type of problem. The equations which govern the action of these analogs are to some scale the same equations which define the action of the prototype. Several examples of this type of computer are aircraft or rocket models in wind tunnels, electric field plotters, and membrane analogy devices. The indirect or mathematical analog makes use of a model—not that of the physical system under consideration, but rather of the equations governing that system. Use is made of various electrical and/or mechanical analogs of the basic mathematical operations: multiplication, addition, integration, etc. The indirect analog is a collection of these "simple" or "subanalogs," so interconnected as to represent a particular equation. Examples of this type are the common slide rule, planimeters, and mechanical and electronic differential analyzers.

Chapter 2 deals specifically with ideas fundamental to the general utilization of electronic digital computers. While the examples contained in this chapter do not, in themselves, describe any particular "real problem," they describe types of operations used in more complex cases. Solutions to specific problem types are contained in later chapters.

It is assumed that the reader is already familiar with a programming language.

2.2 The Computer Program

From the above introduction, it is evident that a set of instructions must be provided to the digital computer which defines in a suitable form the sequence of consecutive operations required to solve a given problem, the solution being the reduction of the problem data to a single number or group of numbers which represents the answer. This detailed set of instructions is called the computer program.

Programs for some of the earlier computing machines were marvels of complexity when compared with present-day standards. Perhaps the best known of these earlier machines was the IBM 650, popular in the late 1950s. These computers were coded or programmed in what is referred to as *machine-language coding*. Only a relative few computer "experts" could do this well, since it involved representing each arithmetic operation, one step at a time, by a series of digits, some describing the operation to be performed and the others indicating where in the computer these results were to be stored. It was necessary that the programmer be familiar not only with the problem being solved but also with the internal workings of the machine.

Because of this severe restriction, and to make the computer available to a larger number of users, there began to appear several problem-oriented programming languages. The purpose of these simplified coding systems was to bridge the wide gap that existed between the user and the computer. Among the more familiar of these languages were Fortran, Mad, and Algol. All these had several things in common. All involved coding systems stated in normal algebraic notation. All required the use of special computer programs referred to as compilers. These compilers served as translators from the user-oriented algebraic programs to their machine-oriented counterparts. Problems coded in one of these languages were therefore run in two phases, or passes: first, the compilation phase; second, the execution phase.

Progress toward more and more sophisticated programming systems has been very rapid over the past several years. This has been made possible in large measure through the great strides that have taken place in the development of faster and more efficient computer hardware. Today, computer systems which can process several programs apparently at the same time are becoming common. These latest generation languages are sometimes referred to as *conversational* in that the computer appears to be almost human in its response to input. It is to be recognized, however, that the language used serves no purpose other than a means of communication between the user and the machine, and therefore can be considered essentially independent of the numerical procedure followed in solving a particular problem.

Problem solution using the digital computer can normally be thought of as being composed of the following four fairly well-defined steps:

NCERN
THIS
OOK

1. Problem formulation
2. Establishment of method of solution
3. Coding
4. Computer executions

Note that the first two of these are independent of the particular machine to be used, and are problem-oriented. The latter two steps, however, depend only on the computer to be used and on its particular coding system. Since it has been observed that most of these systems are easily self-taught, and since there exist many excellent programmed learning texts having to do with this problem, attention will be focused throughout the remainder of this text on the development of logical sequences of operations for the solutions to given problem types. Concern, therefore, will be primarily focused on the "flow" of the computational process.

2.3 *The Flow Diagram*

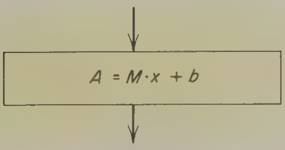
Exactness, clearness, and conciseness are essential features of a computer program. Not only must all contingencies be included, but the basis on which a selection is made must also be indicated.

To minimize the possibility of error and to provide a more desirable form for presentation, there has been developed a more or less standard form of shorthand for use in digital-computer work. This pictorial representation of the logic of a problem solution is called a *flow diagram.*

Several conventions have been developed to represent the various types of operations needed in a solution. In each, a particular geometric form has been selected to represent a specific type of operation. For example, in most, a rectangular box or symbol, as shown in Fig. 2.1, is used to indicate arithmetic operations. The operation in the case illustrated is the computation of the sum of the product Mx and b, followed by the substitution of the sum for the variable called A. It must be understood that the equal-sign used in a flow diagram for any arithmetic operation means that the quantity to the left of the equal-sign is replaced by the quantity to the right. That is, any information that might have been in A is replaced by the new sum. This being the case, $i = i + 1$ is a perfectly valid arithmetic statement and says that the current value of i should be replaced by i plus 1, or simply incremented by 1.

The degree of detail included in a flow diagram is entirely up to the individual programmer. For example, any of the diagrams shown in Fig. 2.2 will accomplish the same objective. The only difference is the

Figure 2.1



[illegible]

OPERATIONS

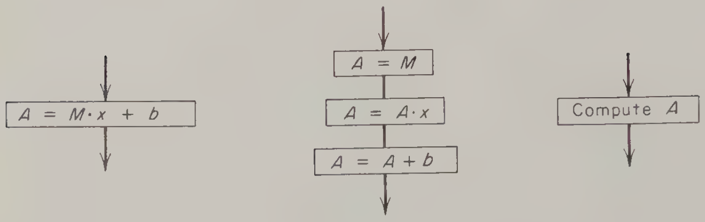


Figure 2.2

amount of detail that has been included. It should be noted that should the last case be used, some other method must be found to indicate the method of determining A.

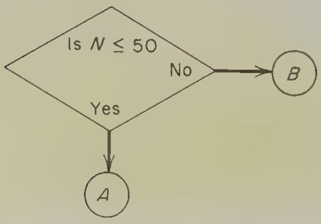
In addition to arithmetic *compute boxes*, symbolic representation for the other needed types of operations must be defined. For example, direction of the flow of data is indicated by an arrow, as shown in Figs. 2.1 and 2.2.

The operation of decision making is represented by a *question box*. Contained therein is a question which must be answerable by either a "yes" or a "no." The alternative exits correspond to the two possibilities. A typical question box is shown in Fig. 2.3. Such a statement requires that the computer check to determine whether the fixed-point variable N is equal to or less than the number 50. If either is true, computation proceeds along the path marked A. If N is greater than 50, the answer to the question is "No," and path B will be followed.

Notes which are to be included in the flow diagram can be contained in circles in the manner shown in Fig. 2.4.

As shown in Fig. 2.5, input operations are represented by a rectangular box with its upper left-hand corner cut off in a fashion similar to that of a punched card; output operations are indicated symbolically by a printed page. The first box stands for the statement "Take in the value assigned to N." Correspondingly, the second case would read "Put out the current values of the variables x, y, and z, in that order."

Figure 2.3



QUESTION
BOX

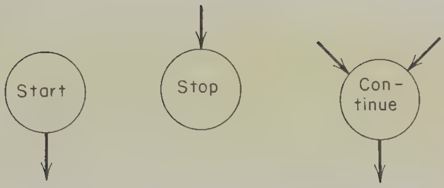


Figure 2.4

To illustrate flow diagramming, and to develop a feeling of confidence in using relatively simple problems, consider the following examples. The chosen problems are not necessarily indicative of engineering-type problems. Rather, they have been selected to demonstrate certain important features of flow diagramming. It is to be understood that in the interest of clarity no attempt has been made to refine the procedures, nor to put them in their most concise or "best" form.

Example 2.1 It is desired to evaluate the quantity $f(x)$ at $x = 2.76$, where

$$f(x) = 2x^2 + 3x + 4 \tag{2.1}$$

Figure 2.6*a* is a flow diagram that could and normally would be used to perform this evaluation. A more detailed flow diagram is given in Fig. 2.6*b*.

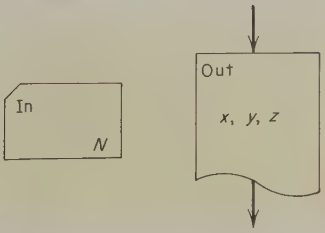
Example 2.2 Evaluate the sum of the products defined by

$$y = \sum_{i=1}^{n} c_i x_i \tag{2.2}$$

In this expression, n is a constant which defines the total number of c's and the total number of x's contained in the problem. The subscript i defines a particular value of c_i and x_i. Figure 2.7 is one possible flow diagram which will perform the summation.

INPUTS

Figure 2.5



OUTPUTS

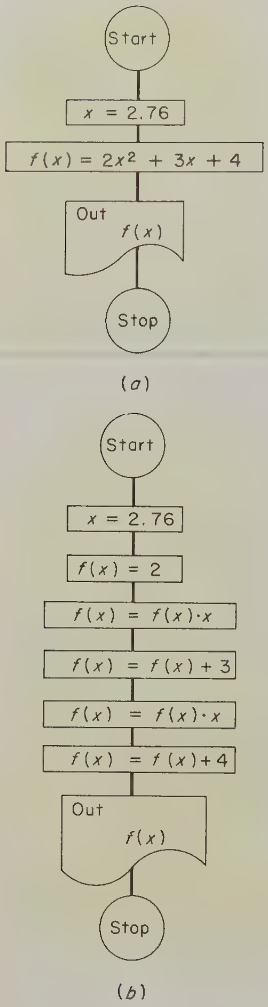


Figure 2.6

The first step in the flow diagram, i.e., the first step in the computer solution to the problem, is to define the value of n, the number of entries of each variable. The values for each of the c_i's and x_i's then are listed. These constitute the input data.

To provide a correct starting point, it is customary and desirable to preset or initialize the dependent variable y. In the case shown, y has been set at an initial value of 0. In addition, the "index" or "counter" i has been initialized to the value 1.

The first computation carried out by the computer is

$$y = 0 + c_1x_1 \tag{2.3}$$

Control then proceeds to the question box, where the current value of i, now 1, is compared with the initially given value of n. If $i = 1$ is not equal to n, exit will be along the path marked "No." The index i will be incremented by 1, and a new value will be computed for y. This second computation is

$$y_{new} = y_{previous} + c_2x_2 \tag{2.4}$$

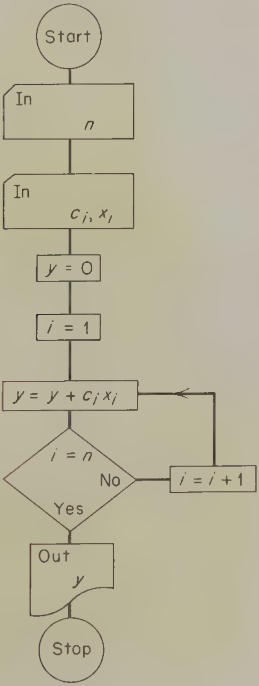
or

$$y = c_1x_1 + c_2x_2 \tag{2.5}$$

Again, i is compared with n and, if required, the process is repeated. When $i = n$, the current value of y is given as output. The computer then stops.

The repetitive nature of this particular example is characteristic of a large number of problem types which effectively utilize the inherent capabilities of a digital computer. The process by which computation was carried out in Example 2.2 is called *looping*. That is, the same basic computational step was repeated over and over, each time with the substitution of a new set of numbers.

Figure 2.7



SUMMATION
PROGRAM

In many cases, it is necessary to loop not just once but twice or even several times. Moreover, loops may be contained one within another, in which case they are said to be "nested." To illustrate this concept, consider Example 2.3.

Example 2.3 Evaluate the double sum defined by

$$\text{Sum} = \sum_{j=1}^{m} \sum_{i=1}^{n} (A_j + 2A_j^2 B_i) \tag{2.6}$$

In this case, both counters i and j must be controlled. The procedure that will be used is as follows: First, set both i and j to their initial values of 1. Then, holding j constant, increment i in steps of 1 from its initial value to n. A new *Sum* will be accumulated at each step. When i reaches n, j will be increased by 1 and i reset to unity. The *increment* i process will be repeated with this new value of j. When j reaches its terminal value m and i equals n, the problem will have been completed.

A flow diagram that will accomplish the desired double summation is shown in Fig. 2.8. Note that, as in Example 2.2, it is necessary to provide the input data of the problem and to initialize to zero the dependent variable *Sum.*

Sets of numbers which can be described in the manner of A_j and B_i of Example 2.3 are said to be *subscripted numbers.* The subscripts are j and i, respectively. Similarly, c_i and x_i of Example 2.2 are subscripted numbers, i being the subscript. In the application of digital computers it is normal procedure to use fixed-point numbers to designate these quantities. This is done primarily for the following two reasons: First, a subscript can take on only integral values. Second, computations involving fixed-point numbers can be carried out more rapidly than those which use floating-point numbers.

A group of numbers the entire set of which can be represented by a single-subscripted variable is called a one-dimensional *array.* The subscript defines the particular entry in question. For example, A_{10} indicates the tenth quantity listed in that group known as A. As will be shown in Chap. 4, a column matrix is an example of a one-dimensional array.

Certain sets of numbers are listed by the designation of two quantities rather than one. When such is the case, a double-subscript notation, such as A_{ij}, can be advantageously used to locate a particular entry. A general matrix (see Chap. 4) is an example of just such a two-dimensional array.

A more or less standard notation has been developed for the definition of two-dimensional arrays. If we assume that it is possible to arrange the

DOUBLE
SUMMATION

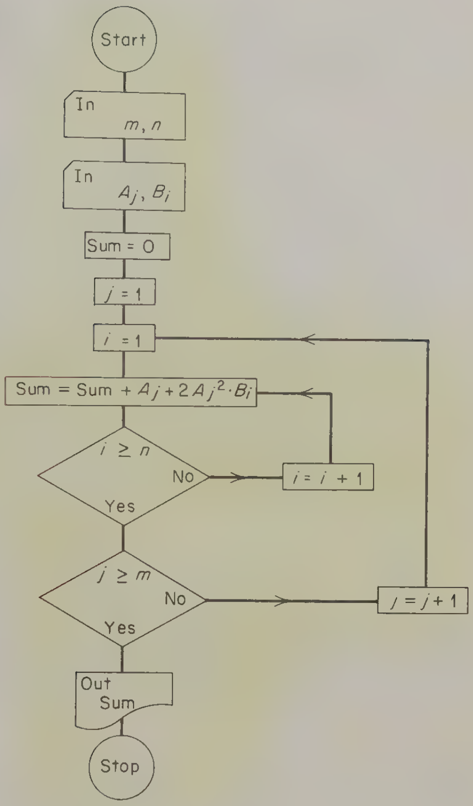
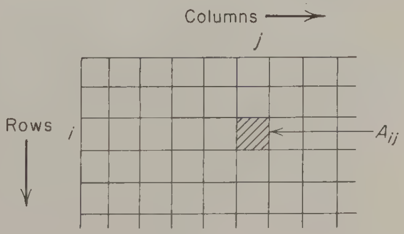


Figure 2.8

quantities in question in a format of rows and columns, A_{ij} is that quantity located in the ith row and the jth column, as shown in Fig. 2.9.

In computations involving two-dimensional arrays, nested loops are not infrequently observed. Example 2.4 is concerned with the manipulation of a square two-dimensional array.

Figure 2.9



Example 2.4 Given A_{ij}, where both i and j range in values from 1 to n, form a new one-dimensional listing called D_i. D_i is to be composed of the diagonal elements of A_{ij}.

The condition defining D_i is

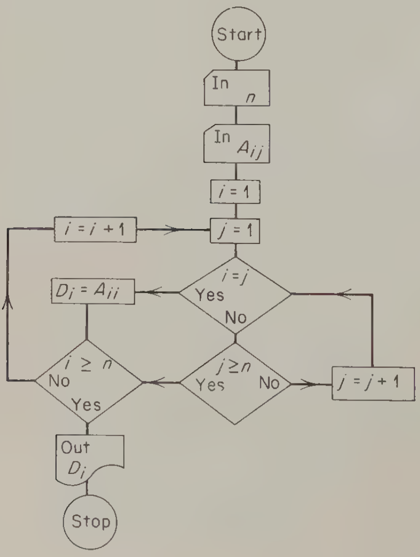
$$D_i = A_{ii} \tag{2.7}$$

The general procedure will be to consider the subscripts i and j and test them for likeness. If, for a particular A_{ij}, the values i and j are found to correspond, that value A_{ij} will be saved as D_i. One possible flow diagram that will solve this problem is given in Fig. 2.10. A more concise flow diagram to accomplish the same purpose is shown in Fig. 2.11.

In the program shown in Fig. 2.10, a large part of the effort is devoted to scanning the array to establish which particular positions correspond to $i = j$. In Fig. 2.11, on the other hand, it is recognized at the outset that the subscripts i and j uniquely locate each entry A_{ij} in the computer storage. Therefore, the proper values can be stipulated and called directly from their stored positions without considering the other elements in the array.

In general, the arithmetic parts of a flow diagram are the easiest to define. More difficult to specify are those segments which deal with the flow of information, branching and looping. For this reason, the remaining examples in this section contain few arithmetic operations beyond sim-

Figure 2.10



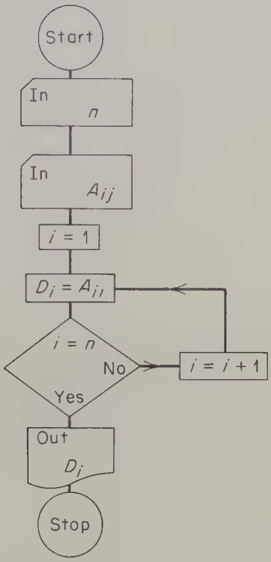


Figure 2.11

ple substitution such as $i = i + 1$. Attention will be focused on the logic of the flow diagram.

Example 2.5 Consider the problem of constructing a flow diagram which will provide as input to a computer a fixed-point number n and n floating-point numbers in the form of a one-dimensional array A_i. When all numbers have been placed in storage, each is to be examined for size and then sorted; the largest number being listed first, the second largest next, and so on until all have been arranged in descending order of magnitude. The computer should then give the final arrangement of the n numbers as output.

The general procedure that will be used to solve this problem will be to select the first given number and compare it with each of the succeeding numbers. Whenever one is found that is larger than the first, that number will be interchanged in position with the first and the process continued. The largest number will correspond to the first in the list after $n - 1$ such comparisons have been made. This number will be saved as A_1. To determine the second largest number, the same process will be repeated, but this time the second number will be used as the base and comparisons will be made with the third, the fourth, and so on to the nth number. The third and subsequent terms will be established in like manner. A flow diagram to accomplish this number sorting is shown in Fig. 2.12.

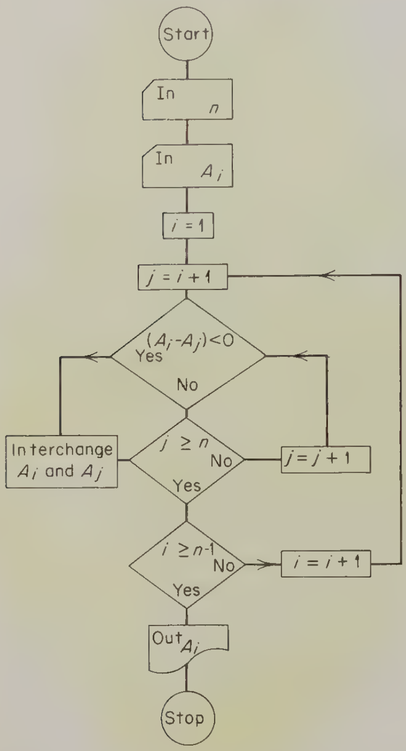


Figure 2.12

In most cases, the best way to examine the working of a flow diagram is to use a numerical example. Consider, for example, the following data and the flow diagram of Fig. 2.12.

$$n = 4 \qquad A_i = \begin{bmatrix} 7 \\ 6 \\ 8 \\ 2 \end{bmatrix} \tag{2.8}$$

First, the value of n and then the n values of A_i are read into the machine. It should be noted that in the diagram two indices i and j have been used. Index i identifies the particular value of A with which all the remaining numbers are to be compared. Index j identifies the remaining numbers. The actual manipulation begins by setting i to 1 and j to $i + 1$, or 2. A check is then made to determine whether A_2 is greater than A_1, that is, whether $A_1 - A_2$ results in a negative quantity. Since $A_1 = 7$ and $A_2 = 6$, it is not the case. Before incrementing j, it is necessary to determine whether all values of A_j have been used. Since

$j = 2$ and not n, j is incremented to 3 and A_1 and A_3 are compared. $A_1 - A_3$ is computed. $A_1 - A_3$ is negative. Thus, A_1 and A_3 are interchanged: A_1 becomes 8, while A_3 becomes 7. Since $j = 3 \neq 4$, j is incremented to 4, and A_1, or 8, is compared with A_4, or 2. The difference $A_1 - A_4$ is found to be positive; and since $j = n$, the largest value of A_i must be the first in the list. At this time the group will be arranged as shown in Eq. (2.9).

$$A_i = \begin{bmatrix} 8 \\ 6 \\ 7 \\ 2 \end{bmatrix} \tag{2.9}$$

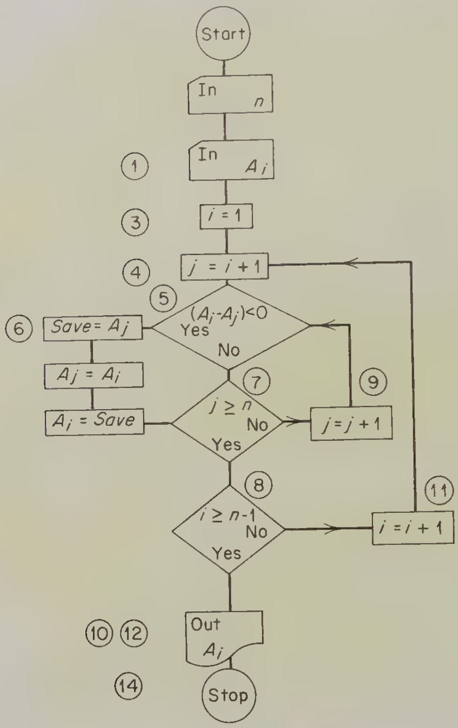
The next step is to determine whether i has become equal to $n - 1$. Since $i = 1$ and $n = 4$, the answer to this test is "no." i is therefore incremented to 2. Correspondingly, j becomes 3. The comparison of A_3 and A_2 requires that the second and third numbers be interchanged. The array now reads

$$A_i = \begin{bmatrix} 8 \\ 7 \\ 6 \\ 2 \end{bmatrix} \tag{2.10}$$

A continuation of the calculational process will show that Eq. (2.10) is the desired array.

Flow diagrams of the type presented in this text define the logic required to solve particular problems. They do not, as was pointed out earlier, constitute the actual itemized instruction that must be supplied to a particular computer. For completeness, then, and to emphasize this point, a Fortran program corresponding to the flow diagram given as Fig. 2.12 is shown in Fig. 2.13. The numbers that are circled in the flow diagram correspond to the statement numbers in the Fortran program.

By comparing the flow diagrams of Figs. 2.12 and 2.13, an inconsistency will be noted. In the first case, one statement defines the process "Interchange A_i and A_j." Three separate arithmetic statements are shown for this same operation in Fig. 2.13. When numbers are interchanged, care must be exercised to ensure that none are lost because of destructive read-in. Figure 2.13 accomplishes this by first defining a new variable, *Save*, into which one of the variables is placed temporarily. The partial flow diagram used is shown in Fig. 2.14. Literally, A_j is placed into a temporary storage location called *Save*. Next, A_i is moved into location A_j. (It should be noted that A_j in that location is destroyed.) Finally, the value that was stored temporarily in *Save* (that is, A_j) is placed in the former A_i location. The interchange is then complete.



```
      DIMENSION  A(1000)
100   FORMAT  (I4)
101   FORMAT  (EI4,8)
      READ  (5,100)N
      M = N - 1
  1   READ(5,101)(A(I),I = 1,N )
  3   I = 1
  4   J = I + 1
  5   IF(A(I) - A(J)) 6,7,7
  6   SAVE = A(J)
      A(J) = A(I)
      A (I) = SAVE
  7   IF(J-N) 9,8,9
  9   J = J + 1
      GO TO  5
  8   IF (I - M) 11,10,11
 11   I = I + 1
      GO TO  4
 10   DO 12 I = 1,N
      WRITE (6,101) A(I)
 12   CONTINUE
 14   CALL  EXIT
      END
```

Figure 2.13

[illegible]
OF DECREASING MAGNITUDE

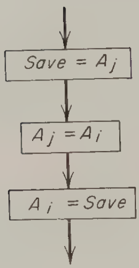


Figure 2.14

Many engineering problems require as part of their solution the scanning of multidimensional arrays. Consider the following example.

Example 2.6 Let A_{ij}, which is a square array of numbers, or a square matrix, of n rows and columns be given.† The problem is to form a new square matrix, the rows of which are the columns of the originally defined

---

† See pp. 40 and 41.

A_{ij}, and vice versa. That is,

$$A_{ij}(new) = A_{ji}(original) \qquad (2.11)$$

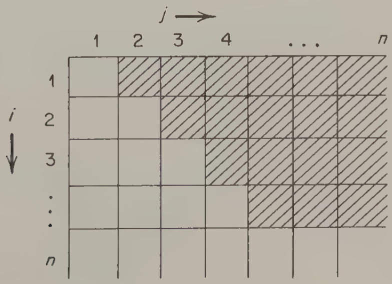
(As will be defined in Chap. 4, the problem is to develop the *transpose* of a given matrix.)

If we assume that it is desirable to use the same variable A_{ij} for the transposed set as for the original set, care must be taken during scanning so as not to destroy half the matrix by destructive read-in. This possibility is minimized if that part of the matrix which lies above or below the main diagonal, which is the diagonal from the upper left or A_{11} element to the lower right or A_{nn} element, is considered first and temporarily stored. For example, consider i ranging from 1 to $n - 1$ and j from $i + 1$ to n. (The region in question is shown shaded in Fig. 2.15.) A flow diagram that will transpose a square matrix is shown in Fig. 2.16.

The purpose of the latter part of Chap. 2 has been to indicate, primarily by example, the basic concepts involved in the construction of flow diagrams. As pointed out, the development of these diagrams is a prerequisite to the preparation of the more detailed digital-computer programs. In many of the illustrative cases, the problems used have little or no direct engineering application. They are included merely to demonstrate a particular technique. Certain of the other examples, however, will be used extensively in the work to follow.

It is the opinion of the authors that before flow diagramming of a particular method of solution to a given problem can be undertaken, the full implications of the proposed method must be understood. Therefore, in each of the following chapters where it is applicable, the development of flow diagrams is delayed to the last section. Only after the theory has been developed, illustrative examples carried out, and problems assigned is any computer formulation considered. Moreover, in most cases, only one of the methods discussed is chosen for flow diagramming.

Figure 2.15



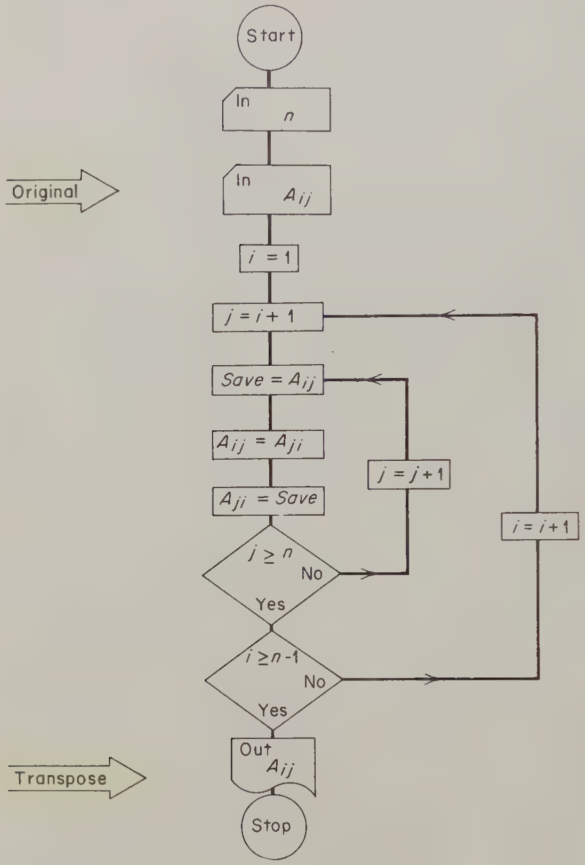


Figure 2.16

Prob. 2.1: Construct a flow diagram that will select the smallest of n ordered numbers A_i, and label that particular number S.

Prob. 2.2: Construct a flow diagram for the evaluation of $n!$.

Prob. 2.3: The average of n numbers is found from the relationship

$$Avg = \frac{1}{n} \sum_{i=1}^{n} A_i$$

Draw a flow diagram to compute Avg.

Prob. 2.4: Develop a flow diagram that will put out the absolute values of a series of numbers A_i.

Prob. 2.5: The numbers M and Q are specified. The quantities Q, $\frac{1}{2}Q$, $\frac{1}{4}Q$, $\frac{1}{8}Q$, etc., are to be cumulatively subtracted from M. Develop a flow diagram that will determine the smallest number of subtractions needed to produce a remainder of less than E. If a solution has not been reached in 100 subtractions, indicate that fact and stop the computation.

Prob. 2.6: Given two linear functions of x,

$$y_1 = f_1(x) \qquad y_2 = f_2(x)$$

develop a flow diagram that will determine the value of x for which

$$|y_2 - y_1| \leq \text{error}$$

Prob. 2.7: Square roots can be obtained with a digital computer by repeated application of the equation

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{N}{x_i}\right)$$

where N = number whose root is desired
x_i = trial value of root
x_{i+1} = computed next trial value of root

Develop a flow diagram in which this process of "iteration" is continued until the difference between x_{i+1} and x_i is equal to or less than a term called *error*.

Hint: In this case, a possible compute box would be

$$\boxed{R = \frac{1}{2}\left(x + \frac{N}{x}\right)}$$

Prob. 2.8: A quadratic equation can be written in the form

$$ax^2 + bx + c = 0$$

If $b^2 - 4ac$ is positive, the equation has two real and unequal roots. These can be found from the relationship

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Construct a flow diagram for determining these roots.

Prob. 2.9: If, in Prob. 2.8, the quantity $b^2 - 4ac$ is negative, the equation has two unequal roots that are complex conjugates. Develop a flow diagram for the determination of these roots.

determinants

3

3.1 Introduction

One of the properties of a square matrix that has particular significance in the solution of a large number of engineering problems is its determinant. Among other things, determinants can be used advantageously in solving sets of linear homogeneous and nonhomogeneous algebraic equations, in determining the nature of the stability of a mechanical or electrical system, and in solving many of the problems of machine theory.

In general, a determinant can be considered as a square array of numbers and/or symbols, and it is usually enclosed by two vertical bars. Two 4 × 4 determinants (or determinants of order 4) are as follows:

$$\begin{vmatrix} 3 & 6 & 2 & -1 \\ 4 & 1 & 6 & 1 \\ 0 & 3 & 1 & 2 \\ -2 & 0 & -2 & 0 \end{vmatrix} \qquad \begin{vmatrix} \sin a & \sin b & \sin c & \sin d \\ a & b & c & d \\ 3 & 1 & 0 & 6 \\ 2a & 4b & c & 3d \end{vmatrix}$$

Often the letter D is used to represent the value of a determinant.

As indicated in Chap. 2, the individual elements in an array are frequently designated by the double-subscript notation a_{ij}. An $n \times n$ determinant would be

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{vmatrix} = |a_{ij}| \tag{3.1}$$

It should be noted that i corresponds to the row and j to the column in which the element in question is located.

Corresponding to every element a_{ij}, there can be written an $n - 1$ square array consisting of the original determinant with the ith row and the jth column removed. This new determinant of order $n - 1$ is called the *minor* of a_{ij} and is designated as M_{ij}. For example, in a determinant of order 3, the minor corresponding to a_{13} is

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \qquad M_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{3.2}$$

Corresponding to each a_{ij} element, there is also a *cofactor*, designated by A_{ij}, where

$$\boxed{A_{ij} = (-1)^{i+j} M_{ij}} \tag{3.3}$$

For example, given the determinant

$$\begin{vmatrix} 3 & 0 & -2 \\ 1 & 4 & 1 \\ 3 & -2 & 5 \end{vmatrix}$$

$$a_{11} = 3 \qquad M_{11} = \begin{vmatrix} 4 & 1 \\ -2 & 5 \end{vmatrix} \qquad A_{11} = (-1)^{1+1} \begin{vmatrix} 4 & 1 \\ -2 & 5 \end{vmatrix}$$

$$a_{32} = -2 \qquad M_{32} = \begin{vmatrix} 3 & -2 \\ 1 & 1 \end{vmatrix} \qquad A_{32} = (-1)^{3+2} \begin{vmatrix} 3 & -2 \\ 1 & 1 \end{vmatrix}$$

3.2 Evaluation of a Determinant

The method for determining the value of a determinant which is most often stressed in high school and elementary college algebra courses is known as *expansion by minors* or *expansion by cofactors*. This method, developed by Laplace, requires the algebraic summation of the n products $a_{ij}A_{ij}$, where either i or j is held fixed while the other takes on the values 1 to n.

Summing along a selected row i:

$$D = \sum_{j=1}^{n} a_{ij}A_{ij} \tag{3.4}$$

Summing along a selected column j:

$$D = \sum_{i=1}^{n} a_{ij}A_{ij}$$

The cofactor for any element in a determinant of order 2 is the diagonally opposite element with the appropriate sign. Thus, the value of a 2×2 determinant is

$$D_{2\times 2} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}(-1)^{1+1}|a_{22}| + a_{12}(-1)^{1+2}|a_{21}|$$

or
$$D_{2\times 2} = a_{11}a_{22} - a_{12}a_{21} \tag{3.5}$$

where use has been made of the fact that the value of a determinant of order 1 is the value of the element in question. Evaluating the preceding example by this method gives the following:

Expansion about the first column:

$$\begin{aligned} D = \begin{vmatrix} 3 & 0 & -2 \\ 1 & 4 & 1 \\ 3 & -2 & 5 \end{vmatrix} &= (3)(-1)^{1+1}\begin{vmatrix} 4 & 1 \\ -2 & 5 \end{vmatrix} + (1)(-1)^{1+2}\begin{vmatrix} 0 & -2 \\ -2 & 5 \end{vmatrix} \\ &\quad + (3)(-1)^{1+3}\begin{vmatrix} 0 & -2 \\ 4 & 1 \end{vmatrix} \\ &= (3)(20+2) - (1)(0-4) + (3)(0+8) = \underline{+94} \end{aligned}$$

Expansion about the first row:

$$\begin{aligned} D &= (3)(-1)^{1+1}\begin{vmatrix} 4 & 1 \\ -2 & 5 \end{vmatrix} + (0)(-1)^{1+2}\begin{vmatrix} 1 & 1 \\ 3 & 5 \end{vmatrix} + (-2)(-1)^{1+3}\begin{vmatrix} 1 & 4 \\ 3 & -2 \end{vmatrix} \\ &= (3)(20+2) + 0 + (-2)(-2-12) = \underline{+94} \end{aligned}$$

It should be noted that the evaluation of an nth-order determinant by the method of expansion by minors requires the evaluation of n determinants of order $n - 1$. These, in turn, each require the evaluation of $n - 1$ determinants of order $n - 2$, etc. The solution of a determinant of order 4 would require the evaluation of

(4)	3×3's
(4)(3)	2×2's
(4)(3)(2)	1×1's

or the evaluation of 4! determinants of order 1. The general $n \times n$ determinant would require $n!$ evaluations. From this it should be clear that the determination of the value of a determinant by the method of expansion by minors (or by cofactors) can be extremely cumbersome and time-consuming.

3.3 *Special Properties of Determinants*

There are certain special properties of determinants that greatly facilitate evaluation. Several of these will be listed here. Their usefulness cannot be overemphasized.

1. A determinant is zero if any two parallel rows (or columns) are identical or multiples one of the other. For example,

$$\begin{vmatrix} 3 & 0 & 6 \\ 1 & 4 & 2 \\ 2 & 1 & 4 \end{vmatrix} = 3(14) + 6(-7) = 0$$

2. A determinant is changed in sign if any two parallel rows (or columns) are interchanged. For example, expanding about the first row,

$$\begin{vmatrix} 3 & 1 & 3 \\ 0 & 2 & 4 \\ 1 & 2 & 3 \end{vmatrix} = -8$$

$$\begin{vmatrix} 3 & 1 & 3 \\ 1 & 2 & 3 \\ 0 & 2 & 4 \end{vmatrix} = +8$$

3. If the transpose of D is defined to be the determinant D' whose columns are the rows of D and vice versa, the value of D' is the same as that of D. For example,

$$D = \begin{vmatrix} 3 & 1 & 3 \\ 0 & 2 & 4 \\ 1 & 2 & 3 \end{vmatrix} = -8$$

$$D' = \begin{vmatrix} 3 & 0 & 1 \\ 1 & 2 & 2 \\ 3 & 4 & 3 \end{vmatrix} = -8$$

4. A common term may be factored from all the elements in a line (row or column) and placed as a multiple outside the new determinant. For example,

$$D = \begin{vmatrix} 2 & 4 & 6 \\ 1 & 0 & 3 \\ 4 & 1 & 2 \end{vmatrix} = 2\begin{vmatrix} 1 & 2 & 3 \\ 1 & 0 & 3 \\ 4 & 1 & 2 \end{vmatrix} = 2(20) = 40$$

5. The value of a determinant is not changed if a multiple of any one of the lines (rows or columns) is added to or subtracted from any parallel line. For example, 3-2 =

4-4=

$$\begin{vmatrix} 3 & 1 & 3 \\ 0 & 2 & 4 \\ 1 & 2 & 3 \end{vmatrix} = \begin{vmatrix} 3 & 1 & \overbrace{3-(2\times 1)} \\ 0 & 2 & 4-(2\times \overset{4}{2}) \\ 1 & 2 & 3-(2\times 2) \end{vmatrix} = \begin{vmatrix} 3 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 2 & -1 \end{vmatrix} = -8$$

3-4 =

3.4 *Row or Column Addition*

Row or column addition, as defined by property 5 above, is particularly important in reducing the amount of work required to evaluate a determinant. Consider the following fourth-order array:

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} \tag{3.6}$$

If we assume that it is desired to reduce all elements in the third column to zero with the exception of the first element, which is to be reduced to 1, a common factor a_{13} is extracted from row 1 (property 4):

$$D = a_{13} \begin{vmatrix} \frac{a_{11}}{a_{13}} & \frac{a_{12}}{a_{13}} & 1 & \frac{a_{14}}{a_{13}} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} \tag{3.7}$$

Row 1 is now multiplied, in turn, by a_{23}, a_{33}, and a_{43}, and the resulting rows are subtracted from each of the remaining rows (property 5). That is,

$$D = a_{13} \begin{vmatrix} \frac{a_{11}}{a_{13}} & \frac{a_{12}}{a_{13}} & 1 & \frac{a_{14}}{a_{13}} \\ a_{21} - \frac{a_{11}}{a_{13}} a_{23} & a_{22} - \frac{a_{12}}{a_{13}} a_{23} & 0 & a_{24} - \frac{a_{14}}{a_{13}} a_{23} \\ a_{31} - \frac{a_{11}}{a_{13}} a_{33} & a_{32} - \frac{a_{12}}{a_{13}} a_{33} & 0 & a_{34} - \frac{a_{14}}{a_{13}} a_{33} \\ a_{41} - \frac{a_{11}}{a_{13}} a_{43} & a_{42} - \frac{a_{12}}{a_{13}} a_{43} & 0 & a_{44} - \frac{a_{14}}{a_{13}} a_{43} \end{vmatrix} \tag{3.8}$$

The order of the determinant can now be reduced by 1 by expansion with the element 1 in the first row, third column as a "pivot."

$$D = a_{13}\begin{vmatrix} a_{21} - \dfrac{a_{11}}{a_{13}}a_{23} & a_{22} - \dfrac{a_{12}}{a_{13}}a_{23} & a_{24} - \dfrac{a_{14}}{a_{13}}a_{23} \\ a_{31} - \dfrac{a_{11}}{a_{13}}a_{33} & a_{32} - \dfrac{a_{12}}{a_{13}}a_{33} & a_{34} - \dfrac{a_{14}}{a_{13}}a_{33} \\ a_{41} - \dfrac{a_{11}}{a_{13}}a_{43} & a_{42} - \dfrac{a_{12}}{a_{13}}a_{43} & a_{44} - \dfrac{a_{14}}{a_{13}}\cancel{a_{44}}\,a_{43} \end{vmatrix} \tag{3.9}$$

By repeated application of this process, any determinant can be reduced to second order, which is evaluated easily. Consider the following example. The line enclosed by dashes is to have all elements reduced to zero save one, the pivotal element, whereas appropriate multiples of the lines underlined by dashes are to be subtracted from parallel lines.

$$\begin{vmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \end{vmatrix} = \begin{vmatrix} 0 & -1 & 0 & -3 \\ 1 & 1 & 1 & 2 \\ 0 & -1 & -1 & -3 \\ 0 & +1 & 0 & 0 \end{vmatrix} = -1\begin{vmatrix} -1 & 0 & -3 \\ -1 & -1 & -3 \\ +1 & 0 & 0 \end{vmatrix}$$

$$= -1\begin{vmatrix} 0 & -3 \\ -1 & -3 \end{vmatrix} = +3$$

3.5 *Pivotal Condensation*

A more formal method for the evaluation of a determinant, which can be derived from the above and which is ideally suited for digital-computer programming, is that of *pivotal condensation.* With the given determinant

$$D = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{3.10}$$

and the assumption that the 1-1 position is to be the pivot point, the order of the determinant can be reduced by 1 by using the following:

$a_{11} \neq 0$

$$D = \frac{1}{(a_{11})^{n-2}}\begin{vmatrix} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{21} & a_{2n} \end{vmatrix} \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{31} & a_{3n} \end{vmatrix} \\ \cdots & \cdots & \cdots & \cdots \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{n1} & a_{n2} \end{vmatrix} & \begin{vmatrix} a_{11} & a_{13} \\ a_{n1} & a_{n3} \end{vmatrix} & \cdots & \begin{vmatrix} a_{11} & a_{1n} \\ a_{n1} & a_{nn} \end{vmatrix} \end{vmatrix} \tag{3.11}$$

Successive application of this general formulation can be used to reduce the order of the determinant to a 2 × 2, which is solved readily.

The example in Sec. 3.4 could be evaluated by the method of pivotal condensation as follows:

$$D = \begin{vmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \end{vmatrix} = \frac{1}{2^2} \begin{vmatrix} 1 & 0 & 3 \\ 0 & -2 & 0 \\ 3 & 0 & 3 \end{vmatrix} = \frac{1}{4}\frac{1}{1} \begin{vmatrix} -2 & 0 \\ 0 & -6 \end{vmatrix} = \frac{12}{4} = +3$$

One of the conditions necessary to obtain a solution by the method of pivotal condensation is that the value of the element a_{11} must be other than zero. If this is not the case, a rearrangement of lines is necessary with a corresponding change in sign of the determinant.

Should a_{11} be a small decimal number, greater accuracy can be achieved by factoring a_{11} out of row 1 or column 1 prior to reduction, that is, by normalizing with respect to a_{11}. The multiplying factor preceding the next-lower-order determinant will then be the "old" value of a_{11} divided by 1.0 raised to the $(n - 2)$nd power, or simply a_{11}.

If certain terms in the array are very large in comparison with those remaining, and if capacity of the computer to be used is relatively small, it may be desirable to factor the largest element from either its row or its column.

Prob. 3.1: Given the relationship $\dfrac{a}{\sin A} = \dfrac{b}{\sin B} = \dfrac{c}{\sin C}$, evaluate the determinant

$$D = \begin{vmatrix} a & x & \sin A \\ b & y & \sin B \\ c & z & \sin C \end{vmatrix}$$

Prob. 3.2: Evaluate by expansion by minors, row or column addition, and pivotal condensation the following determinants:

$$-45 = (a) \begin{vmatrix} 1 & 2 & 0 & 4 \\ 3 & 0 & -1 & -2 \\ 0 & -1 & 3 & 0 \\ 5 & 0 & 1 & 1 \end{vmatrix} \qquad (b) \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \\ 1 & 2 & 3 & 4 \end{vmatrix}$$

$$+5 = (c) \begin{vmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{vmatrix} \qquad (d) \begin{vmatrix} A & -B & C \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} \qquad (e) \begin{vmatrix} 0 & -c_{12} & +c_{13} \\ +c_{12} & 0 & -c_{23} \\ -c_{13} & +c_{23} & 0 \end{vmatrix}$$

Prob. 3.3: By any of the methods discussed, solve the following determinant and arrange the resulting polynomial equation in order of descending powers of e.

$$\begin{vmatrix} e_{11} - e & e_{12} & e_{13} \\ e_{12} & e_{22} - e & e_{23} \\ e_{13} & e_{23} & e_{33} - e \end{vmatrix}$$

Prob. 3.4: Verify, using a general nth-order determinant and a_{11} as a pivot point, that row or column addition and pivotal condensation result in the same $n - 1$ determinant.

Prob. 3.5: Determine the pivotal-condensation reduction equation for expansion of a general nth-order determinant about position a_{23}.

Prob. 3.6: Verify that the value of a determinant having zeros in all elements to the left (or to the right) of the leading diagonal (i.e., the a_{ii} elements) is equal to the product of the diagonal elements.

Prob. 3.7: Determine the positive value of m such that the following determinant is zero:

$$\begin{vmatrix} 1 - m & 2 \\ 2 & 1 - m \end{vmatrix}$$

3.6 Flow Diagram for Pivotal Condensation

In this section a flow diagram will be developed for the evaluation of an nth-order determinant by using pivotal condensation. With the exception of the added normalization step (Step III in Fig. 3.1), the equation governing the solution is that given previously as Eq. (3.11).

The sequence of presentation that will be used in this and the remaining flow-diagramming problems is first to define in very broad terms the major steps that constitute the solution to the problem in question. This general solution then will be followed by detailed discussions of the operations necessary to accomplish each major step. Finally, the entire flow diagrams will be given.

Figure 3.1 defines the major operations to be carried out when evaluating a determinant by pivotal condensation. For reference, each is designated as a step in the solution.

STEP I

Figure 3.2 is a flow diagram that will carry out Step I of Fig. 3.1, the entering of data and initializing of quantities. The notes in Fig. 3.2 have the following significance:

1. n is the order of the determinant to be evaluated.
2. n^2 elements a_{ij} are read into the computer.
3. A special counter k is initialized to 1. Its purpose is to keep track of the number of "pivots," or, alternatively stated, the number of reductions by order 1 of the original nth-order determinant. On

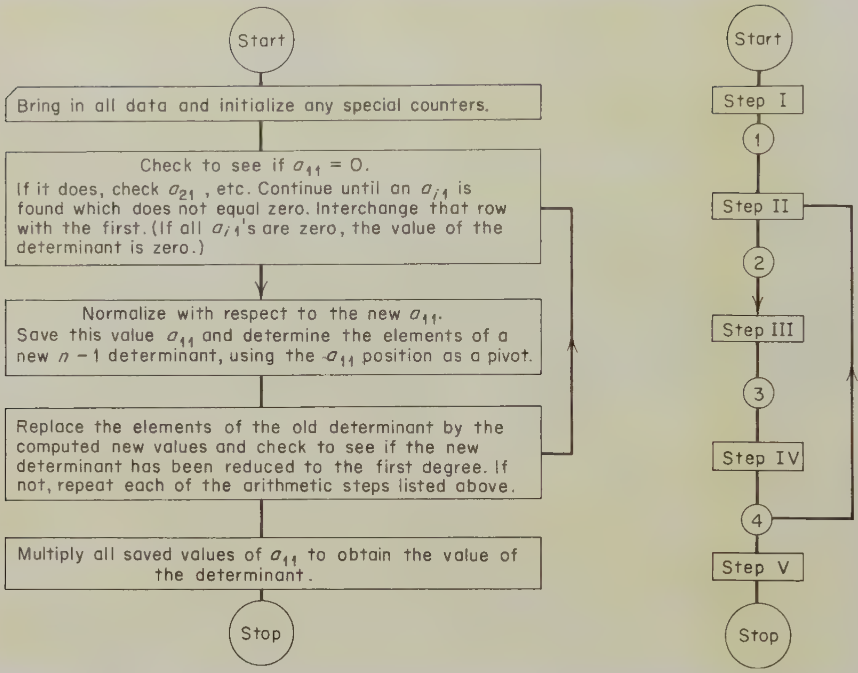
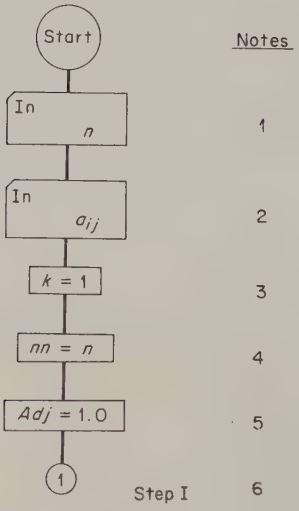

Start
Bring in all data and initialize any special counters.
Check to see if $a_{11} = 0$.
If it does, check a_{21}, etc. Continue until an a_{i1} is found which does not equal zero. Interchange that row with the first. (If all a_{i1}'s are zero, the value of the determinant is zero.)
Normalize with respect to the new a_{11}.
Save this value a_{11} and determine the elements of a new $n-1$ determinant, using the a_{11} position as a pivot.
Replace the elements of the old determinant by the computed new values and check to see if the new determinant has been reduced to the first degree. If not, repeat each of the arithmetic steps listed above.
Multiply all saved values of a_{11} to obtain the value of the determinant.
Stop
Start
Step I
1
Step II
2
Step III
3
Step IV
4
Step V
Stop

Figure 3.1

Figure 3.2



completion of the problem, it should have a value equal to $n - 1$.

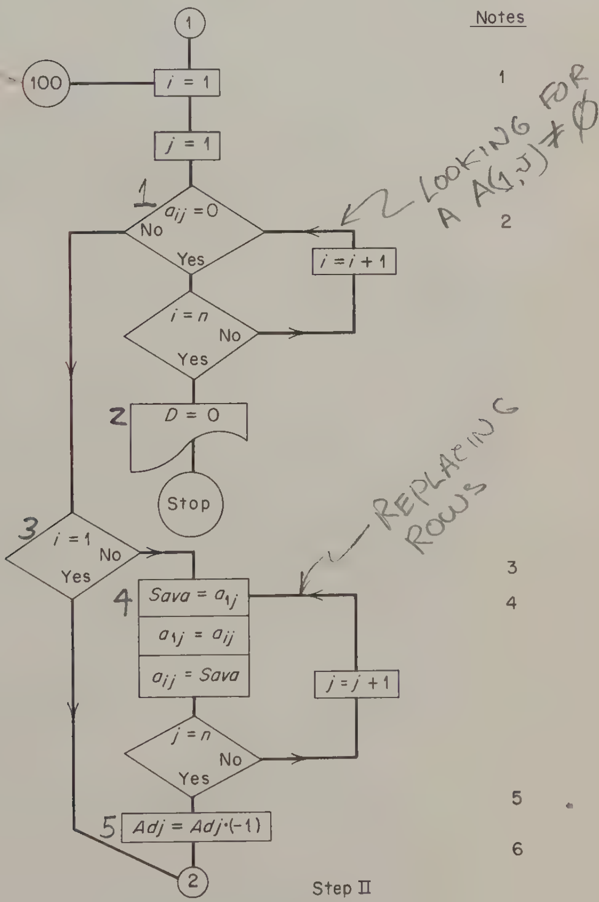
4. nn is a dummy variable set aside to retain the original order of the determinant.
5. Adj is a variable used to compute changes in sign that are introduced into the solution if there is a rearrangement of rows.
6. This is the starting point for Step II of the solution.

STEP II

Check to see if $a_{11} = 0$. If it does, check a_{21}, etc. Continue until an a_{i1} is found which does not equal zero. Interchange that row with the first. (If all a_{i1}'s are zero, the value of the determinant is zero.) With

LOOPING

Figure 3.3



Step II

regard to the notes shown in Fig. 3.3:

1. The values of i and j are both equated to 1, as the first a_{ij} to be considered is a_{11}. The number 100 is a point of entry (for looping) to be used when considering the entire diagram.
2. a_{11} is checked first to determine whether it equals zero. If the check is positive, i is incremented by 1 and a_{21} is examined. This process continues until either i becomes equal to n, at which time the program stops, or a value a_{i1} is found which is not equal to zero.
3. If a_{11} is not equal to zero, the "No" exit of note 2 is followed, and the answer to question 3 is "yes," with control passing to point 2. If, however, $a_{11} = 0$ and an $a_{i1} \neq 0$ is found, row 1 must be interchanged with that particular ith row.
4. The general procedure for interchanging rows was discussed in Example 2.5. A flow diagram was given as Fig. 2.14.
5. If a rearrangement of rows is required, the value of the determinant must be multiplied by the quantity (-1). (See property 2 in Sec. 3.3.) Since *Adj* was initially set equal to 1, the first such rearrangement causes *Adj* to equal -1. Additional switching, at later stages in the solution, will result in other changes in the sign of *Adj*.
6. Control passes from this point to Step III.

STEP III

Normalize with respect to the new a_{11}. Save this value a_{11} and determine the elements of a new $n - 1$ determinant, using the a_{11} position as a pivot.

This part of the flow diagram is shown as Fig. 3.4, with notes as follows:

1. At this point, the current value of a_{11} is placed in $Save_k$. (If $k = 1$, and no row rearrangement was made, $Save_1$ will contain the original a_{11}.) Each reduction by order 1 of the original nth-order determinant produces a new a_{11} element, and thus a new pivot, each of which must be saved. The value of k defines the number of reductions of order 1 that have been made and thus indicates the stage at which a particular a_{11} was calculated.
2. Prior to this stage, j was equal to n. Since a_{1j} is to be divided by a_{11} (that is, quantities are to be normalized with respect to a_{11}), j must be reset to 1. It should be noted that this gives a 1.0 for the pivot value at all times.
3. Again, the values of i and j are reset to 1. By using the 1-1 position as a pivot, a new determinant of order $n - 1$, defined as $Anew_{ij}$, is calculated.
4. i and j at this stage are both equal to $n - 1$. Control passes through point 3 to Step IV.

SAVES THE PIVOTAL VALUE EACH TIME

j MIGHT HAVE BEEN CHANGED IN ABOVE STEPS

CALCULATE VALUES OF ALL THE SMALL DETERMINATES (2×2) AND PUTS THESE VALUES IN A NEW DETERMINATE ANEW ij. A11 = 1 SO IT DOES NOT APPEAR IN THE CALCULATION

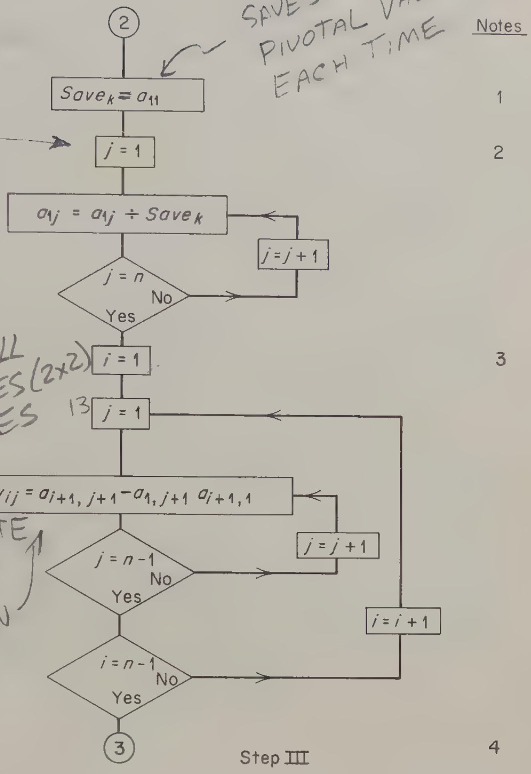


Figure 3.4

$a_{1,1}$ $a_{(i,\,j+1)}$

$a_{(i+1,\,j)}$ $a_{(i+1,\,j+1)}$

STEP IV

Replace the elements of the old determinant by the computed new values and check to see if the new determinant has been reduced to the first degree. If not, repeat each of the arithmetic steps listed above.

In Fig. 3.5, the following are noted:

1. Beginning with i and j equal to 1, the previous values of a_{ij} are replaced, one at a time, by $Anew_{ij}$. The resulting determinant is one order less than the case just preceding.
2. At this stage, a test is required to determine whether n has been reduced to 2. If it has not, n is further reduced by 1 and control is sent back to point 100. The entire process then is repeated. If, however, n is equal to 2, the current determinant is of order $n - 1$, a_{11} is the only element remaining, and control passes to Step V.

SSIGNING THE REDUCED DETERMINATE
NEW (I,J) TO A(I,J) OF REDUCED ORDER

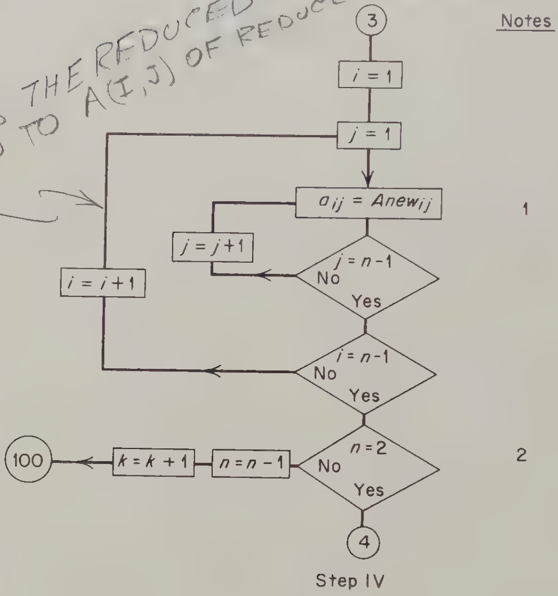
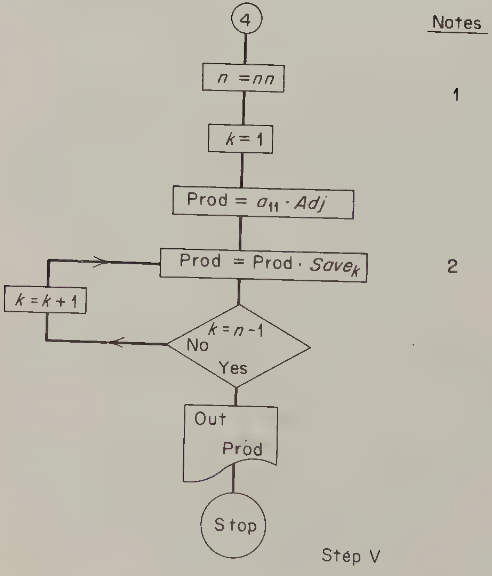


Figure 3.5

Figure 3.6



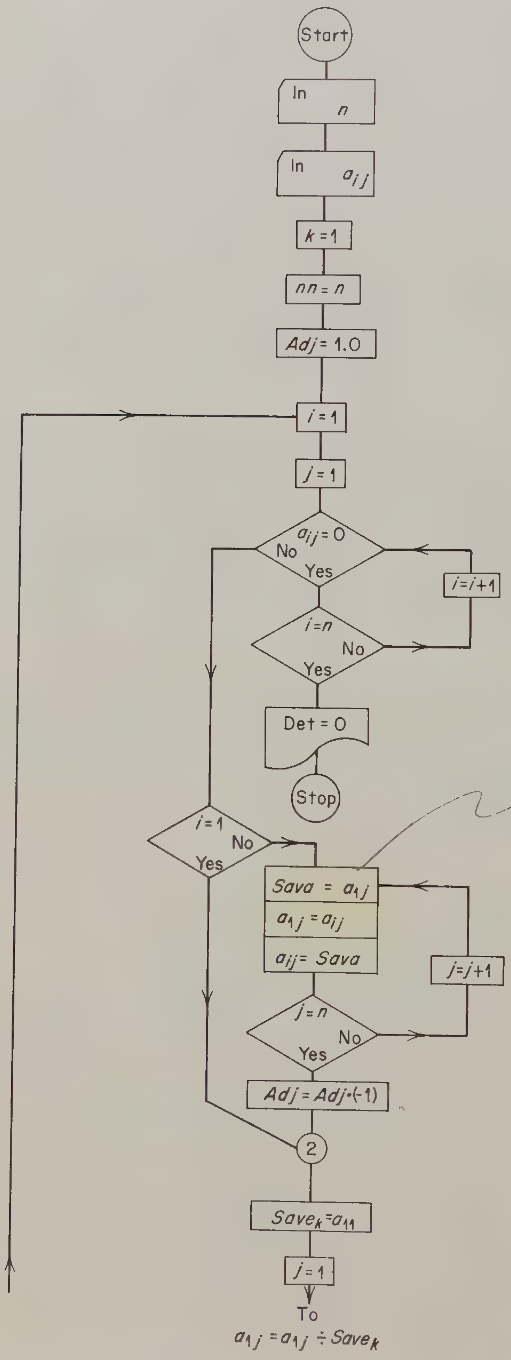
STEP V

Multiply all saved values of a_{11} to obtain the value of the determinant.

In Fig. 3.6, the following are noted:

1. The value of the order of the original determinant is retrieved from its stored location *nn*. The current value of a_{11} is multiplied by the value *Adj* and, depending on the number of row interchanges that have been made during the course of the solution, it can have either a plus or a minus sign.
2. All the previous a_{11}'s are taken from storage ($Save_k$) and in turn multiplied by the product of $a_{11} \cdot Adj$. The final value of the determinant is obtained after $n - 1$ a_{11}'s have been incorporated into the product, the nth a_{11} having been used already in the initial product $a_{11} \cdot Adj$.

Figure 3.7 (pp. 38 and 39) is a complete flow diagram for evaluating an nth-order determinant by pivotal condensation. Again, it should be understood that *this is not a computer program.* It is a diagram, independent of computer make or size, which describes the logic of the solution and indicates the flow of information. It is a necessary step in the writing of a detailed computer program, but it is not the program itself.



ROW
SWITCHING

Figure 3.7 Flow diagram for pivotal condensation.

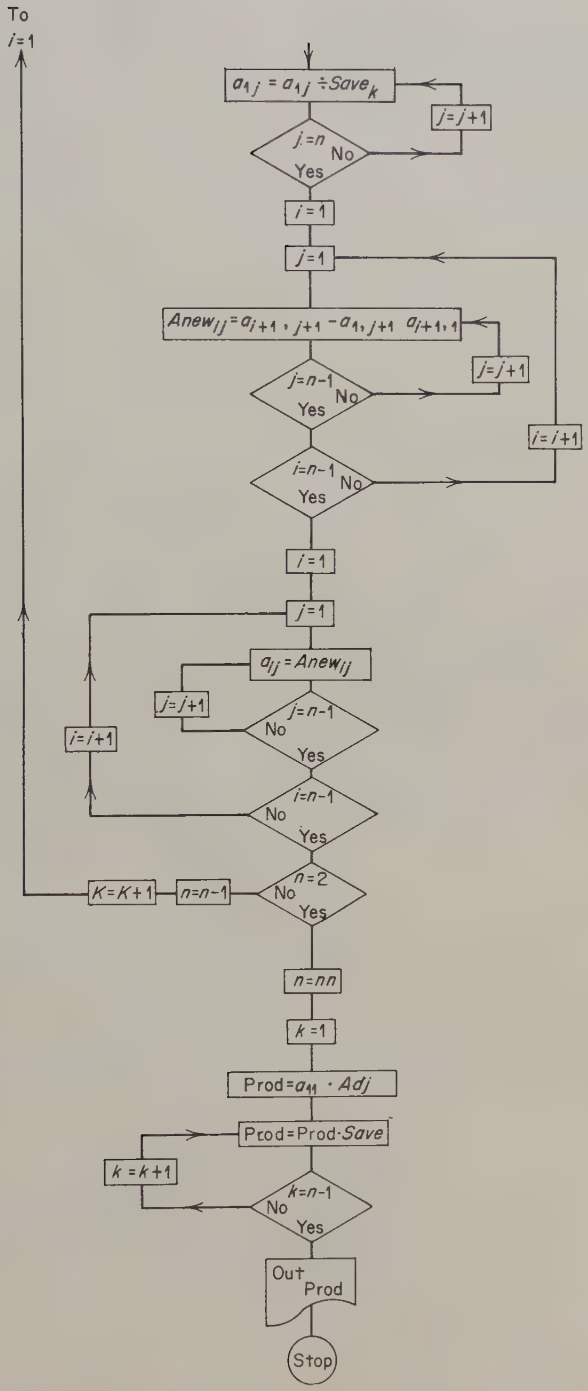


Figure 3.7 *(Continued)*

matrices

4

4.1 Introduction

As was pointed out in Chap. 1, matrix theory offers a convenient means of defining and solving many engineering problems.

To illustrate what is meant by a matrix, suppose that m quantities $y_1, y_2, \ldots, y_m$ are related to n quantities $x_1, x_2, \ldots, x_n$ by the relationship

$$y_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, 2, \ldots, m) \tag{4.1}$$

In expanded form, these equations are

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{aligned} \tag{4.2}$$

By defining a certain ordered way of reading, these also can be written in the following somewhat different form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{4.3}$$

If the column of m entries $y_1, y_2, \ldots, y_m$ is denoted by the single letter $[Y]$, the column of n quantities $x_1, x_2, \ldots, x_n$ by the letter $[X]$, and the m times n quantities a_{ij} by the letter $[A]$, that is,

$$[Y] = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{bmatrix} \qquad [X] = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \qquad [A] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

the original set of equations can be written equally well as

$$[Y] = [A][X] \tag{4.4}$$

The m quantities y_i (that is, $[Y]$) represent an m-row one-column array of numbers or symbols. The n quantities x_j similarly correspond to an n-row one-column group. By definition, a rectangular array of numbers or symbols consisting of one column is called a column matrix or *column vector*, and the individual entries are the components or coordinates of the vector. Correspondingly, an array consisting of a single row is called a *row vector.* The array of coefficients a_{ij} (that is, $[A]$) is called a coefficient matrix or just a matrix. In different forms,

$$[A] = [a_{ij}] \qquad (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n) \tag{4.5}$$

A matrix, then, is nothing more than a rectangular array of numbers and/or symbols with each element being distinct and separate.

By comparing the expanded form of Eqs. (4.2) with the matrix formulation of Eq. (4.3), it is seen that y_i is found by applying the ith row of $[A]$ to the column vector $[X]$ and summing all the products thus obtained. That is,

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \tag{4.6}$$

It will be shown later that this process of row-to-column multiplication and summation is actually the defined process called matrix multiplication.

Matrix formulation has the advantage of brevity. This, however, may not be its greatest virtue. Given the set of equations

$$[Y] = [A][X]$$

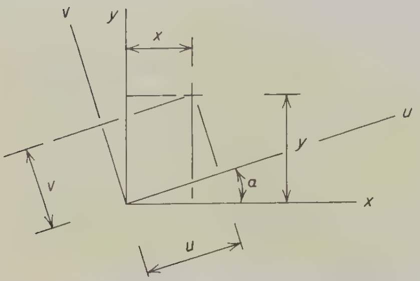


Figure 4.1

or its alternative form

$$y_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, 2, \ldots, m)$$

it is evident that from a mathematical point of view the particular column vectors $[X]$ and $[Y]$ (or x_j and y_i) are not in themselves important. The quantity of significance is the coefficient matrix $[A]$. It is an abstract representation of a linear transformation that exists relating the x's and the y's.

Consider the following equations of transformation from a rectangular xy coordinate system to a rotated, similar uv system having the same origin (see Fig. 4.1). The governing equations are

$$\begin{aligned} u &= x \cos\alpha + y \sin\alpha \\ v &= -x \sin\alpha + y \cos\alpha \end{aligned} \tag{4.7}$$

In matrix notation these equations are written as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4.8}$$

or

$$\begin{bmatrix} u \\ v \end{bmatrix} = [A] \begin{bmatrix} x \\ y \end{bmatrix} \tag{4.9}$$

where $[A]$ is the matrix of the coefficients of x and y and is square (equal number of rows and columns). The elements u and v and x and y constitute column vectors. Assume that it is desired now to effect a similar transformation from uv to a still different system rs by rotation through an additional angle β. The corresponding matrix equation is

$$\begin{bmatrix} r \\ s \end{bmatrix} = [B] \begin{bmatrix} u \\ v \end{bmatrix} \tag{4.10}$$

where

$$[B] = \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \tag{4.11}$$

In matrix form, it is easy to anticipate the result of a multiple transformation from xy to uv and then to rs.

$$\begin{bmatrix} r \\ s \end{bmatrix} = [B]\begin{bmatrix} u \\ v \end{bmatrix} = [B][A]\begin{bmatrix} x \\ y \end{bmatrix} \tag{4.12}$$

It is equally easy to describe a similar multiple transformation in an n-dimensional space.

$$\begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_n \end{bmatrix} = [B]\begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = [B][A]\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{4.13}$$

(It should be noted that in this case the coordinates have been designated as $x_1, x_2, \ldots, x_n$, $u_1, u_2, \ldots, u_n$, and $r_1, r_2, \ldots, r_n$ rather than as different letters for each of the $3n$ coordinate directions.)

4.2 *Special Matrices*

In the work which follows, use will be made of certain special matrices, which are summarized here for definition and future reference.

4.2.1 The Square Matrix

A matrix having an equal number of rows and columns is called a square matrix.

↙ DETERMINATE

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = [a_{ij}] \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, n) \tag{4.14}$$

4.2.2 The Diagonal Matrix

A square matrix in which all elements not on the leading diagonal (that is, the ii position) are zero is called a diagonal matrix.

$$\begin{bmatrix} \alpha_{11} & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & \alpha_{nn} \end{bmatrix} \tag{4.15}$$

4.2.3 The Unit or Identity Matrix

A diagonal matrix in which all the diagonal elements equal 1.0 (that is, $\alpha_{ii} = 1$) is called a unit matrix or an identity matrix. It usually is designated by the letter $[I]$.

$$[I] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{4.16}$$

4.2.4 The Upper Triangular Matrix

A matrix in which zeros exist in all elements to the left of the ii position is called an upper triangular matrix. It frequently is designated by the letter $[U]$.

$$[U] = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \tag{4.17}$$

4.2.5 The Upper Unit Triangular Matrix

An upper triangular matrix having all elements in the leading-diagonal positions equal to 1.0 (that is, $a_{ii} = 1$) is called an upper unit triangular matrix. It frequently is designated by the letter $[T]$, and the individual entries by t_{ij}.

$$[T] = \begin{bmatrix} 1 & t_{12} & \cdots & t_{1n} \\ 0 & 1 & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{4.18}$$

4.2.6 The Lower Triangular Matrix

A matrix in which zeros exist in all elements to the right of the ii position is called a lower triangular matrix. It frequently is denoted by the letter $[L]$ and the entries by ℓ_{ij}.

$$[L] = \begin{bmatrix} \ell_{11} & & \cdots & 0 \\ \ell_{21} & \ell_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{bmatrix} \tag{4.19}$$

4.2.7 The Null or Zero Matrix

A matrix in which zeros exist in all elements is called a null or zero matrix. It is usually designated by the symbol $[0]$ or just $\mathbf{0}$.

4.2.8 The Transpose of a Matrix

The transpose of an $m \times n$ matrix $[A]$ is an $n \times m$ matrix whose columns and rows have been interchanged. It is usually denoted by a prime (that is, $[A']$).

$$[A] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad [A'] = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \tag{4.20}$$

4.3 *Elementary Matrix Operations*

Since by definition a matrix is a rectangular array of numbers, each of which is distinct and separate, two matrices are equal if and only if all their corresponding elements are equal. That is,

$$[A] = [B] \tag{4.21}$$

if $$a_{ij} = b_{ij} \qquad (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$$

4.3.1 Matrix Addition

A matrix $[C]$ is the sum of matrices $[A]$ and $[B]$ if each of the elements in $[C]$ equals the sum of each of the corresponding elements in $[A]$ and $[B]$. That is,

$$c_{ij} = a_{ij} + b_{ij} \tag{4.22}$$

Consider the following equation:

$$\sum_{j=1}^{n} a_{ij}x_j - \sum_{j=1}^{n} b_{ij}x_j = 0 \qquad (i = 1, 2, \ldots, n) \tag{4.23}$$

or $$\sum_{j=1}^{n} (a_{ij} - b_{ij})x_j = 0 \qquad (i = 1, 2, \ldots, n) \tag{4.24}$$

or, in matrix forms,

$$[A][X] - [B][X] = [[A] - [B]][X] = [0] \tag{4.25}$$

or

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = [0] \tag{4.26}$$

Matrix addition requires that

$$[C][X] = [A - B][X] = [0] \tag{4.27}$$

In expanded form, this is

$$\begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1n} - b_{1n} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2n} - b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} - b_{n1} & a_{n2} - b_{n2} & \cdots & a_{nn} - b_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = [0] \tag{4.28}$$

4.3.2 Matrix Multiplication

To illustrate the general concept of matrix multiplication, consider the following set of equations in x_i:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= f_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= f_2 \end{aligned} \qquad \text{or} \qquad [A][X] = [F] \tag{4.29}$$

Assume, now, that x_1, x_2, and x_3 are defined in terms of two other variables y_1 and y_2.

$$\begin{aligned} x_1 &= b_{11}y_1 + b_{12}y_2 \\ x_2 &= b_{21}y_1 + b_{22}y_2 \\ x_3 &= b_{31}y_1 + b_{32}y_2 \end{aligned} \qquad \text{that is,} \qquad [B][Y] = [X] \tag{4.30}$$

It therefore would be possible to write the original equations (4.29) in terms of the two variables y_1 and y_2.

$$\begin{aligned} c_{11}y_1 + c_{12}y_2 &= f_1 \\ c_{21}y_1 + c_{22}y_2 &= f_2 \end{aligned} \qquad \text{or} \qquad [C][Y] = [F] \tag{4.31}$$

By substituting from Eqs. (4.30) the values of the x's into Eqs. (4.29), the matrix corresponding to $[C]$ of Eqs. (4.31) is

$$[C] = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \tag{4.32}$$

As suggested earlier in this section, the total process could be defined equally well as

$$[A][B] = [C] \tag{4.33}$$

or

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \tag{4.34}$$

It should be noted that matrix multiplication consists in row-to-column multiplication and summation. For example, the 1-1 position in matrix $[C]$ is obtained by multiplying and summing the elements of row 1 of matrix $[A]$ to those of column 1 of matrix $[B]$. Similarly, c_{12} is found by multiplying and summing row 1 of matrix $[A]$ with column 2 of matrix $[B]$.

Having considered the special case of multiplying a 2 × 3 matrix by a 3 × 2 matrix, consider now the general statement of the matrix multiplication problem. Assume that the following two transformations are given:

$$[Y] = [A][X] \qquad \text{and} \qquad [X] = [B][Z] \tag{4.35}$$

In summation form, these are

$$y_i = \sum_{j=1}^{n} a_{ij}x_j \qquad (i = 1, 2, \ldots, m) \tag{4.36}$$

and

$$x_j = \sum_{k=1}^{q} b_{jk}z_k \qquad (j = 1, 2, \ldots, n) \tag{4.37}$$

y_i in terms of z_k is then

$$y_i = \sum_{j=1}^{n} a_{ij} \sum_{k=1}^{q} b_{jk}z_k = \sum_{k=1}^{q} \sum_{j=1}^{n} a_{ij}b_{jk}z_k \tag{4.38}$$

Since it is the composite effect of the two transformations that is being sought,

$$[Y] = [C][Z] \tag{4.39}$$

or

$$y_i = \sum_{k=1}^{q} c_{ik}z_k \qquad (i = 1, 2, \ldots, m) \tag{4.40}$$

A comparison of Eqs. (4.38) and (4.40) yields

$$c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk} \qquad (i = 1, 2, \ldots, m;\ k = 1, 2, \ldots, q) \tag{4.41}$$

Matrix multiplication $[C] = [A][B]$ requires that the elements of row i and column k of the matrix $[C]$ equal the sum of the pairs of products of the elements of row i of $[A]$ and column k of $[B]$.

It should be noted that $[Y] = [A][X]$, where $[X]$ is a column vector, is a true matrix multiplication.

4.4 Special Considerations Concerning Matrix Multiplication

From the development of Eqs. (4.41) it is evident that the matrix multiplication $[C] = [A][B]$ has meaning only when the number of columns of $[A]$ equals the number of rows of $[B]$.

4.4.1 Multiplication by the Identity Matrix

The multiplication of any matrix by an identity matrix yields the original matrix.

$$[A] = [I][A] = [A][I] \tag{4.42}$$

or

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{4.43}$$

4.4.2 Matrix Multiplication Is Associative

$$[A][B\ C] = [A\ B][C] \tag{4.44}$$

If $[A] = a_{ij}$, $[B] = b_{jk}$, and $[C] = c_{k\ell}$, the ik position of the product $[A][B] = [A\ B] \equiv [D]$ is [from Eqs. (4.41)]

$$d_{ik} = \sum_j a_{ij} b_{jk} \tag{4.45}$$

and the $i\ell$ position of $[A\ B][C]$ is

$$\sum_k d_{ik} c_{k\ell} = \sum_k \sum_j a_{ij} b_{jk} c_{k\ell} \tag{4.46}$$

Similarly, the product $[A][B\ C] = [A][E]$ is (for the $i\ell$ position)

$$\sum_j a_{ij} e_{j\ell} = \sum_j \sum_k a_{ij} b_{jk} c_{k\ell} \tag{4.47}$$

All that differs in Eqs. (4.46) and (4.47) is the order of summation. Matrix multiplication therefore is associative.

4.4.3 Matrix Multiplication Is Not in General Commutative ($[A\ B] \neq [B\ A]$)

To illustrate this condition, consider the multiplication of two 2 × 2 matrices $[A]$ and $[B]$.

$$[A\ B] = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \tag{4.48}$$

and

$$[B\ A] = \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} \end{bmatrix} \tag{4.49}$$

From Eqs. (4.48) and (4.49), it is evident that only for certain special values of the a_{ij}'s and b_{ij}'s will the two be equal.

4.4.4 Matrix Multiplication Is Distributive

$$[A][B+C] = [A][B] + [A][C] \tag{4.50}$$

(*Note:* matrix addition.) If $[A] = a_{ij}$, $[B] = b_{jk}$, and $[C] = c_{jk}$, then from Eq. (4.23)

$$[B+C] = b_{jk} + c_{jk} \tag{4.51}$$

The entry in the ik position for the product $[A][B+C]$ is

$$\sum_j a_{ij}(b_{jk} + c_{jk}) = \sum_j a_{ij}b_{jk} + \sum_j a_{ij}c_{jk} \tag{4.52}$$

From Eq. (4.21) it is also evident that Eq. (4.52) also represents the matrix sum of the two multiplications

$$[A][B] + [A][C] \tag{4.53}$$

The distributive condition therefore is demonstrated.

4.4.5 Multiplication by a Constant

The multiplication of any matrix by a constant (scalar) results in a new matrix, every entry of which is multiplied by the constant.

$$\alpha \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} \alpha & 0 & \cdots & 0 \\ 0 & \alpha & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha a_{n1} & \alpha a_{n2} & \cdots & \alpha a_{nn} \end{bmatrix} \tag{4.54}$$

4.5 *Partitioning of Matrices*

It is evident that as matrices become larger, the amount of work required to carry out a given operation (e.g., multiplication) may increase at a fast rate. Computer storage capacity may also become a problem in solving the larger problems. To facilitate solution, it is often desirable to partition matrices into subgroups or submatrices and to operate on these smaller units.

Consider the 4 × 4 matrix

$$[A] = \left[\begin{array}{ccc:c} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ \hdashline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right] \tag{4.55}$$

For the divisions indicated by the dotted lines, there results a matrix of order 3, a 1 × 3 row vector, a 3 × 1 column vector, and a number a_{44}. This can therefore be written as shown in Eq. (4.56).

$$[A] = \left[\begin{array}{ccc|c} & & & a_{14} \\ & [A_{3\times3}] & & a_{24} \\ & & & a_{34} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right] \tag{4.56}$$

or

$$[A] = \begin{bmatrix} [A_{3\times3}] & [H_{3\times1}] \\ [G_{1\times3}] & a_{44} \end{bmatrix} \tag{4.57}$$

where the column vector (a_{14},a_{24},a_{34}) is denoted by $[H_{3\times1}]$ and the row vector (a_{41},a_{42},a_{43}) by $[G_{1\times3}]$. This type of partitioning, known as *bordering*, has particular advantage in the solution of certain types of engineering problems.

A special array, which is "naturally" partitional and which is frequently observed, consists of square subsets lying along the leading diagonal with all other terms equal to zero.

$$\left[\begin{array}{cc|ccc} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ \hline 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & a_{43} & a_{44} & a_{45} \\ 0 & 0 & a_{53} & a_{54} & a_{55} \end{array}\right] = \begin{bmatrix} [A_{11}] & 0 \\ 0 & [A_{22}] \end{bmatrix} \tag{4.58}$$

where

$$[A_{11}] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{and} \quad [A_{22}] = \begin{bmatrix} a_{33} & a_{34} & a_{35} \\ a_{43} & a_{44} & a_{45} \\ a_{53} & a_{54} & a_{55} \end{bmatrix} \tag{4.59}$$

To illustrate multiplication on a partitioned matrix, consider the following:

$$[A][B] = [C] \tag{4.60}$$

where $[A]$ is a 2 × 3 and $[B]$ is a 3 × 2 matrix. From Eq. (4.34), the transformation yields

$$\begin{aligned} [C] &= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \\ &= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \end{aligned} \tag{4.61}$$

Given the same sets partitioned as shown by the dotted lines,

$$\begin{aligned} [C] &= \left[\begin{array}{c|cc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{array}\right] \left[\begin{array}{cc} b_{11} & b_{12} \\ \hline b_{21} & b_{22} \\ b_{31} & b_{32} \end{array}\right] = [[A_1] \quad [A_2]] \begin{bmatrix} [B_1] \\ [B_2] \end{bmatrix} \\ &= [A_1 \quad B_1] + [A_2 \quad B_2] \end{aligned} \tag{4.62}$$

It should be noted that the last of these is a matrix addition problem. The solution in expanded form is then

$$[C] = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{21}b_{11} & a_{21}b_{12} \end{bmatrix} + \begin{bmatrix} a_{12}b_{21} + a_{13}b_{31} & a_{12}b_{22} + a_{13}b_{32} \\ a_{22}b_{21} + a_{23}b_{31} & a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \tag{4.63}$$

or $$[C] = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \tag{4.64}$$

which is the same as Eq. (4.61).

In general, if a matrix is partitioned into submatrices, the diagonals of which are all equal in order (that is, all submatrices are square and of the same order), thereby ensuring that all elementary matrix operations can be realized, the cells (i.e., the submatrices) can be operated on as if they were individual elements of a matrix. For example, consider the following two matrices $[A]$ and $[B]$ in which the individual elements $[A_{ij}]$ and $[B_{ij}]$ are assumed to be of equal order:

$$[A] = \begin{bmatrix} [A_{11}] & [A_{12}] & \cdots & [A_{1n}] \\ [A_{21}] & [A_{22}] & \cdots & [A_{2n}] \\ \cdots & \cdots & \cdots & \cdots \\ [A_{n1}] & [A_{n2}] & \cdots & [A_{nn}] \end{bmatrix}$$

and $$[B] = \begin{bmatrix} [B_{11}] & [B_{12}] & \cdots & [B_{1n}] \\ [B_{21}] & [B_{22}] & \cdots & [B_{2n}] \\ \cdots & \cdots & \cdots & \cdots \\ [B_{n1}] & [B_{n2}] & \cdots & [B_{nn}] \end{bmatrix} \tag{4.65}$$

The matrix sum is

$$[A + B] = \begin{bmatrix} [A_{11} + B_{11}] & [A_{12} + B_{12}] & \cdots & [A_{1n} + B_{1n}] \\ [A_{21} + B_{21}] & [A_{22} + B_{22}] & \cdots & [A_{2n} + B_{2n}] \\ \cdots & \cdots & \cdots & \cdots \\ [A_{n1} + B_{n1}] & [A_{n2} + B_{n2}] & \cdots & [A_{nn} + B_{nn}] \end{bmatrix} \tag{4.66}$$

Matrix multiplication requires that

$$[AB] = \begin{bmatrix} [C_{11}] & [C_{12}] & \cdots & [C_{1n}] \\ [C_{21}] & [C_{22}] & \cdots & [C_{2n}] \\ \cdots & \cdots & \cdots & \cdots \\ [C_{n1}] & [C_{n2}] & \cdots & [C_{nn}] \end{bmatrix} \tag{4.67}$$

where

$$[C_{ij}] = [A_{i1}][B_{1j}] + [A_{i2}][B_{2j}] + \cdots + [A_{in}][B_{nj}] \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, n) \tag{4.68}$$

For those cases where $[A_{ij}]$ and $[B_{ij}]$ are not necessarily square or equal in order, the required condition for partitioning is that the products and/or sums be realized. [Consider, for example, the bordered case of Eq. (4.62).]

To illustrate the matrix formulation of engineering problems, consider the following examples.

Example 4.1 Given the ordinary differential equation

$$\frac{d^4y}{dx^4} + k^2 \frac{d^2y}{dx^2} = 0 \tag{4.69}$$

subjected to the boundary conditions

$$\begin{aligned} &\text{At } x = 0: \\ &\qquad y = 1 \qquad y' = 0 \\ &\text{At } x = 1: \\ &\qquad y = 0 \qquad y'' = 0 \end{aligned} \tag{4.70}$$

in which

$$y' = \frac{dy}{dx} \qquad \text{and} \qquad y'' = \frac{d^2y}{dx^2}$$

The general solution to Eq. (4.69) is

$$y = A \sin kx + B \cos kx + Cx + D \tag{4.71}$$

The problem is to define the constants A, B, C, and D. By substituting the boundary conditions into the general solution, the following four equations in the unknown constants are obtained:

$$\begin{aligned} &\text{At } x = 0,\ y = 1: \\ &\qquad A(0) + B(1) + C(0) + D(1) = 1 \\ &\text{At } x = 0,\ y' = 0: \\ &\qquad A(k) + B(0) + C(1) + D(0) = 0 \\ &\text{At } x = 1,\ y = 0: \\ &\qquad A(\sin k) + B(\cos k) + C(1) + D(1) = 0 \\ &\text{At } x = 1,\ y'' = 0: \\ &\qquad A(-k^2 \sin k) + B(-k^2 \cos k) + C(0) + D(0) = 0 \end{aligned} \tag{4.72}$$

In matrix form these are

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ k & 0 & 1 & 0 \\ \sin k & \cos k & 1 & 1 \\ \sin k & \cos k & 0 & 0 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4.73}$$

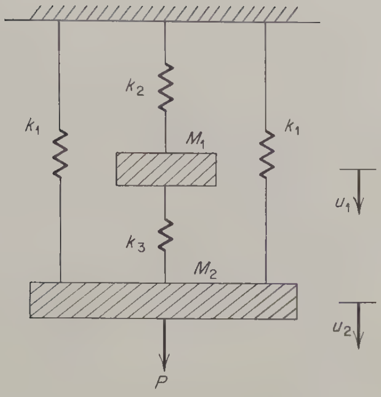


Figure 4.2

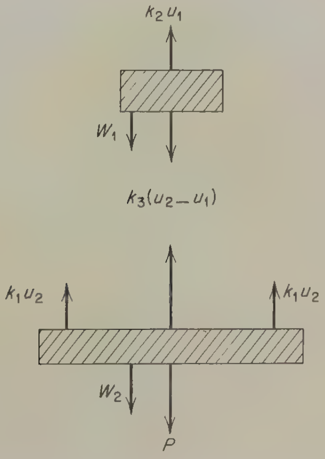
Example 4.2 The two-mass spring-connected system shown in Fig. 4.2 is subjected to a constant force P. The problem is to formulate the steady-state expressions relating the displacements u_1 and u_2, the imposed force P, the spring constants k_1, k_2, and k_3, and the masses M_1 and M_2.

By considering the force acting on each of the masses M_1 and M_2 as represented by individual free-body diagrams and summing forces in the vertical direction (see Fig. 4.3), Eqs. (4.74) and (4.75) define the problem.

$$k_2u_1 = k_3(u_2 - u_1) + W_1 \tag{4.74}$$

$$W_2 + P = 2k_1u_2 + k_3(u_2 - u_1) \tag{4.75}$$

Figure 4.3



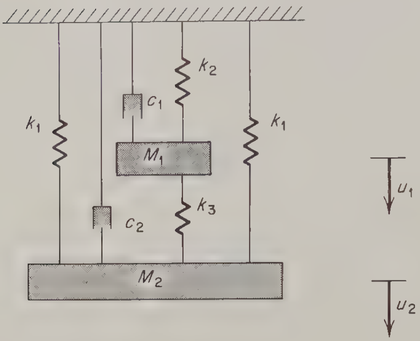


Figure 4.4

In matrix form, these are

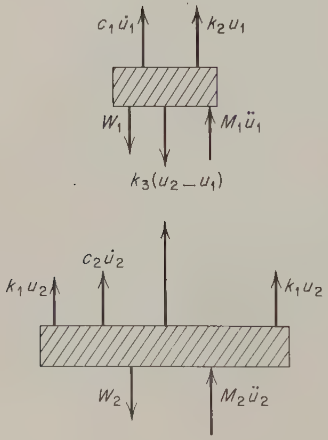
$$\begin{bmatrix} k_2 + k_3 & -k_3 \\ -k_3 & 2k_1 + k_3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} W_1 \\ P + W_2 \end{bmatrix} \tag{4.76}$$

NOT Resp.

Example 4.3 To illustrate matrix formulation in transient-type problems, assume that the same mass-spring system of Fig. 4.2 is depressed a certain amount vertically and then released. Also, assume that a proportional to velocity damping is present in the system (see Fig. 4.4).

At the instant the system is released from its disturbed state, the free-body diagrams with d'Alembert forces are those shown in Fig. 4.5.

Figure 4.5



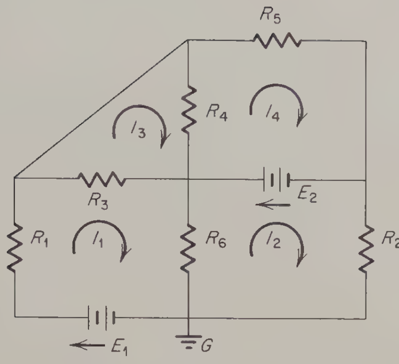


Figure 4.6

Again summing vertical forces,

$$k_2u_1 - k_3(u_2 - u_1) + M_1\ddot{u}_1 + c_1\dot{u}_1 = W_1 \tag{4.77}$$

$$2k_1u_2 + k_3(u_2 - u_1) + M_2\ddot{u}_2 + c_2\dot{u}_2 = W_2 \tag{4.78}$$

These can be written as

$$[A]\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + [B]\begin{bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \end{bmatrix} + [C]\begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \tag{4.79}$$

where $$[A] = \begin{bmatrix} k_2 + k_3 & -k_3 \\ -k_3 & 2k_1 + k_3 \end{bmatrix} \qquad [B] = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix} \tag{4.80a}$$

and $$[C] = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix} \tag{4.80b}$$

It should be noted that the coefficient matrix of the u's, that is, the matrix $[A]$, is identical with that of Example 4.2. It is a property of the spring system and not of the imposed loading.

R,

Example 4.4 A dc network is illustrated in Fig. 4.6. Imposed voltages are E_1 and E_2 (directions indicated), and the resistances are R_1, $R_2, \ldots, R_6$. By assuming the unknown loop currents I_1, I_2, I_3, and I_4 and using Kirchhoff's law—that the net voltage drop in any closed loop must equal zero—the following equations can be written:

Loop 1: $$E_1 - R_1I_1 - R_3(I_1 - I_3) - R_6(I_1 - I_2) = 0 \tag{4.81}$$

Loop 2: $$-E_2 - R_2I_2 - R_6(I_2 - I_1) = 0 \tag{4.82}$$

Loop 3: $$-R_3(I_3 - I_1) - R_4(I_3 - I_4) = 0 \tag{4.83}$$

Loop 4: $$E_2 - R_4(I_4 - I_3) - R_5(I_4) = 0 \tag{4.84}$$

In matrix form, these equations can be written as

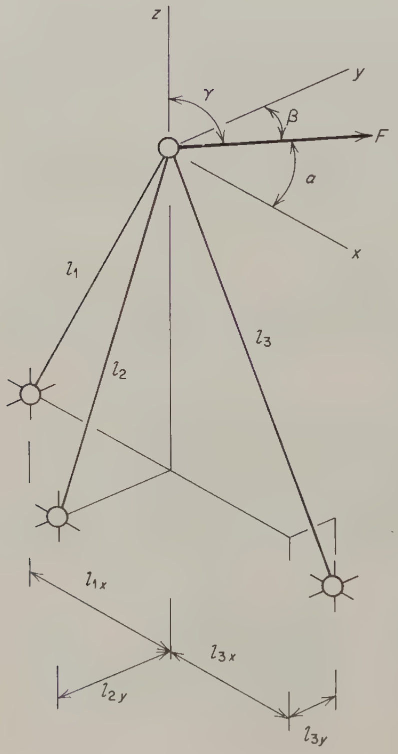
$$\begin{bmatrix} R_1 + R_3 + R_6 & -R_6 & -R_3 & 0 \\ -R_6 & R_2 + R_6 & 0 & 0 \\ -R_3 & 0 & R_3 + R_4 & -R_4 \\ 0 & 0 & -R_4 & R_4 + R_5 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} = \begin{bmatrix} E_1 \\ -E_2 \\ 0 \\ E_2 \end{bmatrix} \tag{4.85}$$

Note that the coefficient matrix is symmetrical about the leading diagonal. Many engineering problems have equations like (4.85).

NR.

Example 4.5 The three-bar pin-connected space truss shown in Fig. 4.7 is subjected to an applied force F. The bar lengths are ℓ_1, ℓ_2, and ℓ_3.

Figure 4.7



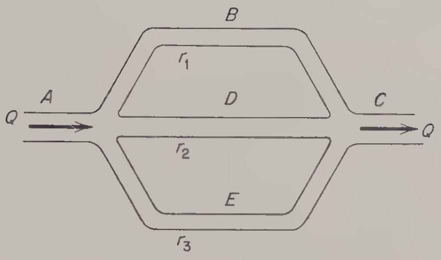


Figure 4.8

Their components in the coordinate directions are listed by the subscripts as shown. For example, the component of ℓ_1 in the x direction is ℓ_{1x}. The applied force F has direction cosines α, β, and γ with the coordinates. The unknown forces in each of the members are F_1, F_2, and F_3. (*Note:* Tension has been assumed as positive.)

$$\begin{aligned}
\Sigma F_x &= 0 \qquad F\cos\alpha - F_1\left(\frac{\ell_{1x}}{\ell_1}\right) + F_3\left(\frac{\ell_{3x}}{\ell_3}\right) = 0 \\
\Sigma F_y &= 0 \qquad F\cos\beta - F_2\left(\frac{\ell_{2y}}{\ell_2}\right) + F_3\left(\frac{\ell_{3y}}{\ell_3}\right) = 0 \\
\Sigma F_z &= 0 \qquad F\cos\gamma - F_1\left(\frac{\ell_{1z}}{\ell_1}\right) - F_2\left(\frac{\ell_{2z}}{\ell_2}\right) - F_3\left(\frac{\ell_{3z}}{\ell_3}\right) = 0
\end{aligned} \tag{4.86}$$

In matrix form, these are

$$\begin{bmatrix} \dfrac{\ell_{1x}}{\ell_1} & 0 & -\dfrac{\ell_{3x}}{\ell_3} \\ 0 & \dfrac{\ell_{2y}}{\ell_2} & -\dfrac{\ell_{3y}}{\ell_3} \\ \dfrac{\ell_{1z}}{\ell_1} & \dfrac{\ell_{2z}}{\ell_2} & \dfrac{\ell_{3z}}{\ell_3} \end{bmatrix} \begin{bmatrix} \dfrac{F_1}{F} \\ \dfrac{F_2}{F} \\ \dfrac{F_3}{F} \end{bmatrix} = \begin{bmatrix} \cos\alpha \\ \cos\beta \\ \cos\gamma \end{bmatrix} \tag{4.87}$$

Note that all quantities have been nondimensionalized.

NR

Example 4.6 To illustrate matrix formulation in a nonlinear system, consider the hydraulic network shown in Fig. 4.8. It is assumed that the fluid is incompressible, that the total rate of flow in at A and out at C is Q, that the individual rates of flow in B, D, and E are q_1, q_2, and q_3, and that the pipe resistances are r_1, r_2, and r_3. It is also presupposed that the pressure drop in the direction of flow is equal to the resistance of the pipe times the square of the rate of flow. That is,

$$\Delta p = r_i q_i^2 \tag{4.88}$$

If the rate of flow in pipe ABC is recalled as q_1, and that in ADC as q_2, the rate of flow in AEC will be

$$q_3 = Q - q_1 - q_2 \tag{4.89}$$

If the pressure at C is assumed to be zero (i.e., discharge into the air), the governing equations will be

ABC versus ADC:

$$r_1 q_1^2 = r_2 q_2^2 \tag{4.90}$$

ADC versus AEC:

$$r_2 q_2^2 = r_3 (Q - q_1 - q_2)^2 \tag{4.91}$$

or

$$q_1^2(r_1) + q_2^2(-r_2) = 0 \tag{4.92}$$

$$q_1^2(r_3) + q_1 q_2 (2r_3) + q_2^2(r_3 - r_2) + q_1 Q(-2r_3) + q_2 Q(-2r_3) = -Q^2 r_3 \tag{4.93}$$

By nondimensionalizing with respect to Q and defining

$$x = \frac{q_1}{Q} \tag{4.94}$$

and

$$y = \frac{q_2}{Q} \tag{4.95}$$

the matrix formulation will be

$$\begin{bmatrix} r_1 & 0 & -r_2 & 0 & 0 \\ r_3 & 2r_3 & r_3 - r_2 & -2r_3 & -2r_3 \end{bmatrix} \begin{bmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ -r_3 \end{bmatrix} \tag{4.96}$$

N. Resp.

Example 4.7 Consider the simultaneous differential equations

$$\begin{aligned} \frac{dx}{dz} + 2x - y &= 3z \\ \frac{dy}{dz} - 3x + y &= e^{2z} \end{aligned} \tag{4.97}$$

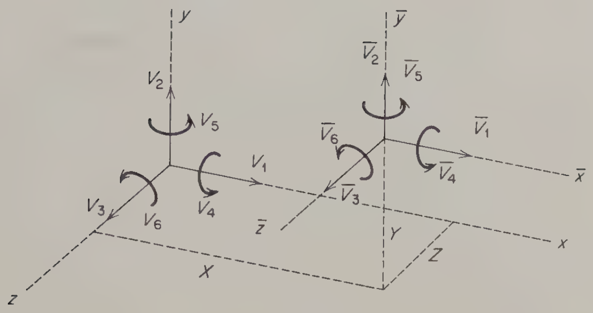


Figure 4.9

Writing these in operational form,

$$
\begin{aligned}
D(x) + 2x - y &= 3z \\
D(y) - 3x + y &= e^{2z}
\end{aligned}
\tag{4.98}
$$

The matrix formulation is

$$
\begin{bmatrix} D+2 & -1 \\ -3 & D+1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3z \\ e^{2z} \end{bmatrix}
\tag{4.99}
$$

NOT Reqd.

Example 4.8 To illustrate a case where the resulting matrix is readily partitional, consider the transformation of a generalized force-couple system in rectangular coordinates x, y, and z to a parallel axis system $\bar{x}$, $\bar{y}$, and $\bar{z}$, which is displaced X, Y, and Z from the original system (see Fig. 4.9). For the two to be equivalent, it is necessary that the force components and moments about each of the axes be equal. That is

$$
\begin{aligned}
\bar{V}_1 &= V_1 & \qquad \bar{V}_4 &= V_4 + ZV_2 - YV_3 \\
\bar{V}_2 &= V_2 & \qquad \bar{V}_5 &= V_5 + XV_3 - ZV_1 \\
\bar{V}_3 &= V_3 & \qquad \bar{V}_6 &= V_6 + YV_1 - XV_2
\end{aligned}
\tag{4.100}
$$

In matrix form

$$
\begin{bmatrix} \bar{V}_1 \\ \bar{V}_2 \\ \bar{V}_3 \\ \bar{V}_4 \\ \bar{V}_5 \\ \bar{V}_6 \end{bmatrix}
=
\left[\begin{array}{ccc:ccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\hdashline
0 & +Z & -Y & 1 & 0 & 0 \\
-Z & 0 & +X & 0 & 1 & 0 \\
+Y & -X & 0 & 0 & 0 & 1
\end{array}\right]
\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \end{bmatrix}
\tag{4.101}
$$

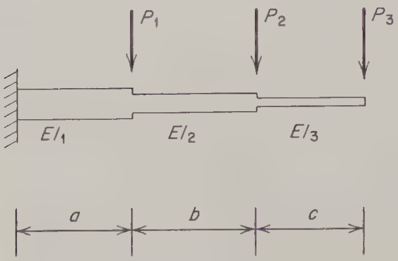


Figure 4.10

or

$$[\bar{V}] = \left[\begin{array}{c|c} [I] & [0] \\ \hline [A] & [I] \end{array}\right] [V] \tag{4.102}$$

where

$$[A] = \begin{bmatrix} 0 & Z & -Y \\ -Z & 0 & X \\ Y & -X & 0 \end{bmatrix} \tag{4.103}$$

Prob. 4.1: Carry out the matrix multiplications indicated in the following:

(*a*) $\begin{bmatrix} \alpha & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$ (*b*) $\begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$

(*c*) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \alpha \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$ (*d*) $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$

(*e*) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$ (*f*) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{bmatrix}$

Prob. 4.2: A 4 × 4 unit upper triangular matrix is multiplied by a 4 × 4 lower triangular matrix; that is, $[L][T] = [C]$. Determine the elements of the general matrix $[C]$.

Prob. 4.3: A cantilever beam of three segments is subjected to any one or any combination of vertical forces P_1, P_2, and P_3. (See Fig. 4.10.) Defining the deflection at the load points in the direction of the loads as δ_i, where

$$\delta_i = \sum_j P_j \delta_{ij} \qquad (i, j = 1, 2, 3)$$

and δ_{ij} is the deflection of point i due to a unit load at point j, determine the matrix equation relating the deflections at points 1, 2, and 3 to the imposed loading.

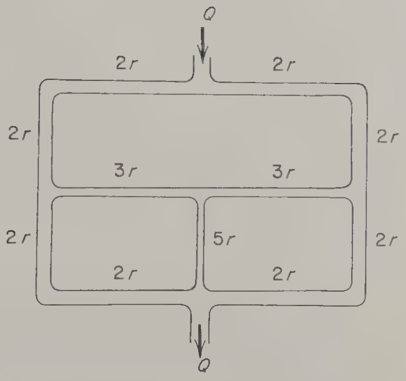


Figure 4.11

Prob. 4.4: For the x, y, and z components of force and moment shown in Fig. 4.9, verify that the transformation equations corresponding to a rotation of the axis system about the original origin is

$$\begin{bmatrix} \bar{V}_1 \\ \bar{V}_2 \\ \bar{V}_3 \\ \bar{V}_4 \\ \bar{V}_5 \\ \bar{V}_6 \end{bmatrix} = \left[\begin{array}{ccc|ccc} a_{\bar{x}x} & a_{\bar{x}y} & a_{\bar{x}z} & 0 & 0 & 0 \\ a_{\bar{y}x} & a_{\bar{y}y} & a_{\bar{y}z} & 0 & 0 & 0 \\ a_{\bar{z}x} & a_{\bar{z}y} & a_{\bar{z}z} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & a_{\bar{x}x} & a_{\bar{x}y} & a_{\bar{x}z} \\ 0 & 0 & 0 & a_{\bar{y}x} & a_{\bar{y}y} & a_{\bar{y}z} \\ 0 & 0 & 0 & a_{\bar{z}x} & a_{\bar{z}y} & a_{\bar{z}z} \end{array}\right] \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \end{bmatrix} = \left[\begin{array}{c|c} [R] & [0] \\ \hline [0] & [R] \end{array}\right] [V] \qquad (4.104)$$

where $a_{\bar{j}i}$ is the direction cosine of the angle between the new $\bar{j}$ direction and the original i axis.

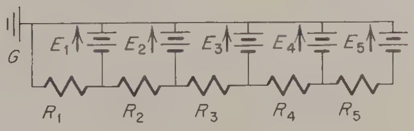
Prob. 4.5: Using Eqs. (4.103) and (4.104), write the matrix equations corresponding to a general rotation and translation of a rectangular-coordinate system.

Prob. 4.6: For the pipe network system shown in Fig. 4.11, assuming incompressible flow, discharge into the air, and a pressure drop according to the relationship

$$\Delta p = r_i q_i^2$$

formulate the matrix equation relating flow in each pipe to the indicated resistances. (Note symmetry.)

Figure 4.12



Prob. 4.7: If unknown loop currents I_i^c in each of the "cells" of Fig. 4.12 are assumed is it possible to relate these currents to the voltages by the matrix equation

$$[A][I^c] = [I][E]$$

where $[A]$ is a function only of the resistances $R_1, R_2, \ldots, R_5$? (*Note:* $[I]$ is the identity matrix.)

Prob. 4.8: Write in matrix form the following equations or expressions:

(*a*) $x_{ab} = \sum_i \sum_j \ell_{ia}\ell_{jb}y_{ij} \qquad (a, b = 1, 2, 3)$

(*b*) Sum of two vectors **A** and **B**

(*c*) Dot product of two vectors **A** and **B** (that is, **A** · **B**)

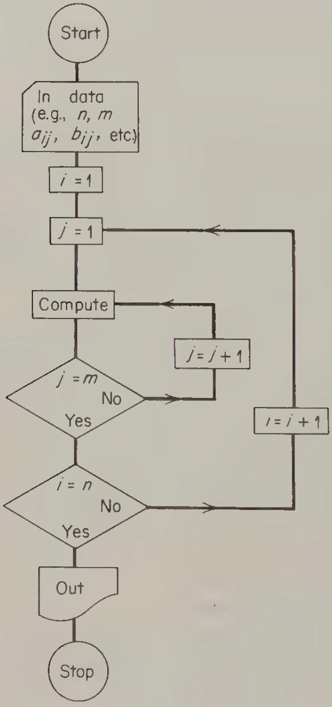
(*d*) Cross product of two vectors **A** and **B** (that is, **A** × **B**)

(*e*) Curl of vector **A**

4.6 Flow Diagrams for Elementary Matrix Operations

It is evident from the illustrative examples of this chapter that matrix algebra is a powerful tool in the formulation and solution of engineering

Figure 4.13



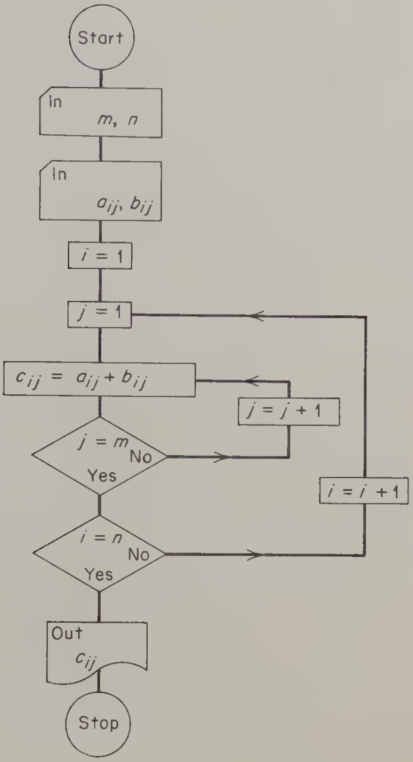


Figure 4.14 Flow diagram for matrix addition.

problems. In matrix form, regardless of the area of engineering from which the problem arises, the uniformity of the basic mathematical problem is most apparent.

Examples 2.4 and 2.6 of the pivotal-condensation problem of Sec. 3.6 were concerned with the digital-computer formulation of specific matrix manipulations. As in those illustrations, flow diagrams for matrix problems invariably require at least one set of nested loops. That is, during the course of the computation, each of the subscripts must be incremented from its initial to its final value. Because situations of this type are observed frequently, either as the total problem or as a segment of the problem, it is considered desirable for emphasis to diagram here again a general procedure for carrying out such processes (see Fig. 4.13).

4.6.1 Matrix Addition

From Eq. (4.23), the defined matrix-addition operation requires that

$$c_{ij} = a_{ij} + b_{ij} \tag{4.105}$$

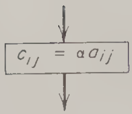


Figure 4.15

Both a_{ij} and b_{ij} must be of equal "size." The new matrix c_{ij} will be the same. A possible flow diagram for matrix addition is shown in Fig. 4.14. It should be noted that the only difference between Figs. 4.13 and 4.14 is the statement contained in the "compute box."

4.6.2 Multiplication of a Matrix by a Constant

The multiplication of a matrix by a constant requires that each and every element in the matrix be multiplied by the constant. [See Eq. (4.54).] With the exception of the instruction in the compute box, the flow diagram for this operation is identical with that for matrix addition (Fig. 4.14). For this second case, the statement shown in Fig. 4.15 should be used as the compute box.

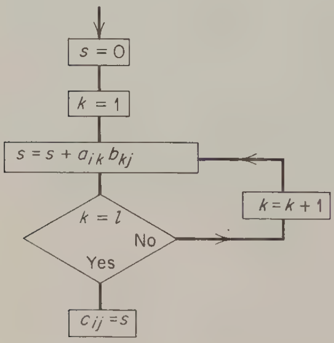
4.6.3 Matrix Multiplication

From Eq. (4.41), the product of two matrices $[A]$ and $[B]$ is given by the expression

$$c_{ij} = \sum_{k=1}^{l} a_{ik}b_{kj} \tag{4.106}$$

It should be noted that three subscripts have been used. This has been done to facilitate flow diagramming. It also emphasizes the fact

Figure 4.16



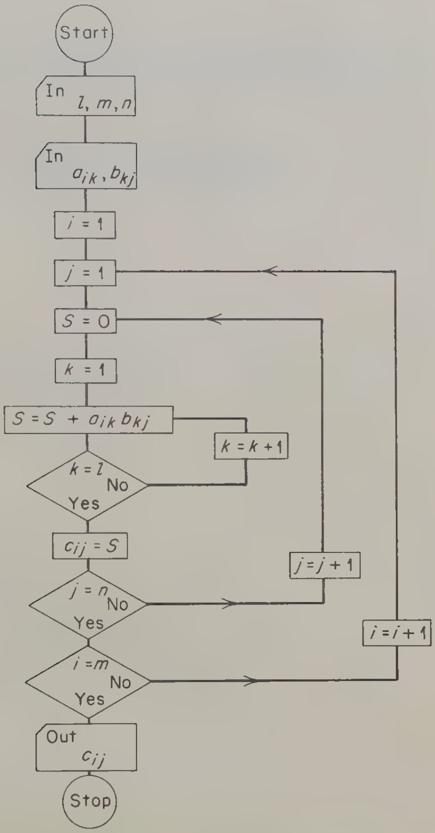


Figure 4.17 Flow diagram for matrix multiplication.

that the number of columns of matrix $[A]$ must equal the number of rows of matrix $[B]$.

Replacing the compute box for this matrix problem is the series of operations shown in Fig. 4.16, where the index k has been incremented from 1 to ℓ. The total flow diagram for matrix multiplication is shown in Fig. 4.17.

5 simultaneous linear algebraic equations

5.1 *Introduction*

An equation of one variable can be represented in functional form as

$$f(x) = c \tag{5.1}$$

The corresponding two-variable problem in x_1 and x_2 is

$$f_1(x_1,x_2) = c_1 \qquad f_2(x_1,x_2) = c_2 \tag{5.2}$$

By extending this system, an n-dimensional set can be written as shown by Eqs. (5.3).

$$\begin{aligned} f_1(x_1,x_2, \ . \ . \ . \ ,x_n) &= c_1 \\ f_2(x_1,x_2, \ . \ . \ . \ ,x_n) &= c_2 \\ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ &. \\ f_n(x_1,x_2, \ . \ . \ . \ ,x_n) &= c_n \end{aligned} \tag{5.3}$$

In the matrix form of Chap. 4, these are

$$[A][X] = [C] \tag{5.4}$$

A simultaneous set is considered to be *nonlinear* if any one of the functions f_i is nonlinear. Conversely, a *linear system* requires that all functions f_i be linear.

The solution of systems of simultaneous linear algebraic equations requires the determination of the particular values of the variables $x_1, x_2, \ldots, x_n$ that will satisfy simultaneously the given equations. This, however, assumes that such values can be obtained. Consider, for example, the following exceptions:

1.

$$\text{DEPENDENT}\left\{\begin{aligned} 6x_1 - 3x_2 &= 6 \\ 2x_1 - x_2 &= 2 \end{aligned}\right. \tag{5.5}$$

By assuming a value of $x_1 = 1.0$, it is observed that the value $x_2 = 0$ will satisfy both equations. For $x_1 = 2.0$, the value of x_2 is found to be $+2.0$. An infinite number of other possibilities also could be found.

x_1	x_2
0	−2.0
+1.0	0
+2.0	+2.0
+3.0	+4.0
.	.
.	.
.	.

(5.6)

2.

$$\text{CONFLICTING}\left\{\begin{aligned} 6x_1 - 3x_2 &= 6 \\ 2x_1 - x_2 &= 6 \end{aligned}\right. \tag{5.7}$$

No solution can be found that will simultaneously satisfy these (conflicting) equations (5.7).

It should be noted that in each of the cases demonstrated above, the value of the determinant of the coefficients of the x_i's equals zero.

$$\begin{vmatrix} 6 & -3 \\ 2 & -1 \end{vmatrix} = 0 \tag{5.8}$$

No unique solutions to the given equations can be found when such is the case.

Certain linear systems behave as if the unknowns (assumed to be independent) are almost dependent. A small variation in the coefficients markedly influences the solution to the problem. Roundoff errors also have a profound effect. Such systems are frequently referred to as *ill-*

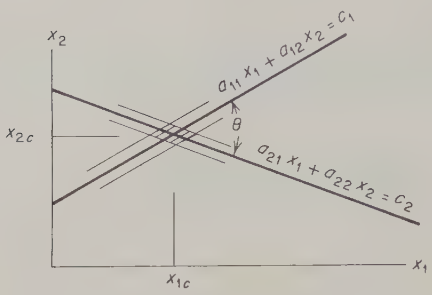


Figure 5.1

conditioned systems. Consider the following two simultaneous equations:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= c_1 \\ a_{21}x_1 + a_{22}x_2 &= c_2 \end{aligned} \tag{5.9}$$

Assume, now, that the a_{ij}'s and c_i's are not absolutely correct but are listed only to a certain accuracy. The true expressions therefore should be

$$\begin{aligned} (a_{11} + \epsilon_{11})x_1 + (a_{12} + \epsilon_{12})x_2 - c_1 &= \gamma_1 \\ (a_{21} + \epsilon_{21})x_1 + (a_{22} + \epsilon_{22})x_2 - c_2 &= \gamma_2 \end{aligned} \tag{5.10}$$

where the ϵ's and γ's represent error terms.

The original equations (5.9) can be represented graphically by the solid lines of Fig. 5.1, where the correct solution is given as x_{1c}, x_{2c}. But γ_1 and γ_2 other than zero result in parallel displacements of the two curves, as indicated by the dotted lines. For ϵ's equal to zero and a prescribed accuracy of the c's defined by the γ's, the solution will lie within the shaded region. The influence of the ϵ's is to change the slopes of the lines and thus further alter the region of possible solutions.

Should the two equations correspond to lines having a small difference in slope—that is, should θ be small—a slight variation in one of the unknowns may result in a significant change in the other. Such systems are considered to be ill conditioned. It should be noted, however, that no fixed angle or function of the angle can be stated that separates an ill-conditioned from a well-behaved set.

The above reasoning can be extended to n-dimensional situations where the region of possible solutions is defined by a series of planes. Here, it is even more difficult to establish limits of permissible variation.

For most of the work that follows, it will be assumed that the systems of equations considered are well behaved. The exception to this will be the latter part of the discussion of iterative methods of solution.

5.2 *Cramer's Rule*

The set of equations under consideration can be represented in summation form as follows:

$$\sum_{j=1}^{n} a_{ij}x_j = c_i \qquad (i = 1, 2, \ldots, n) \tag{5.11}$$

By premultiplying both sides of this equation by a summed set of arbitrary (at this time) cofactors A_{ik}, Eq. (5.11) becomes

$$\sum_{i=1}^{n} A_{ik} \sum_{j=1}^{n} a_{ij}x_j = \sum_{i=1}^{n} c_i A_{ik} \tag{5.12}$$

Since the order in which the summation is carried out is not critical, this can be written as

$$\sum_{j=1}^{n} x_j \sum_{i=1}^{n} a_{ij}A_{ik} = \sum_{i=1}^{n} c_i A_{ik} \tag{5.13}$$

But since the product of an element and any cofactor other than its own is zero,

$$\sum_{i=1}^{n} a_{ij}A_{ik} \tag{5.14}$$

will have values other than zero only when $k = j$. Equation (5.13) can therefore be written as

$$x_j \sum_{i=1}^{n} a_{ij}A_{ij} = \sum_{i=1}^{n} c_i A_{ij} \qquad (j = 1, 2, \ldots, n) \tag{5.15}$$

If the process of expansion by cofactors (or minors) is recalled from Chap. 3, this can be written as

$$x_j D = D_j \qquad (j = 1, 2, \ldots, n) \tag{5.16}$$

or

$$\boxed{x_j = \frac{D_j}{D}} \qquad (j = 1, 2, \ldots, n) \tag{5.17}$$

where D is the value of the determinant of the coefficients a_{ij}, and D_j is the value of the same determinant with the exception that the jth column of the a_{ij}'s has been replaced by the column vector $c_1, c_2, \ldots, c_n$. Equation (5.17), known as *Cramer's rule*, can be written in expanded form as

$$x_j = \frac{\begin{vmatrix} a_{11} & a_{12} & \cdots & c_1 & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & c_2 & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & c_n & \cdots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nj} & \cdots & a_{nn} \end{vmatrix}} \tag{5.18}$$

To illustrate the solution of a problem by this method, consider the following set of simultaneous algebraic equations:

$$\begin{bmatrix} +3 & -3 & +5 \\ +1 & +2 & -6 \\ +2 & -1 & +3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +4 \\ +3 \\ +1 \end{bmatrix} \tag{5.19}$$

For this case

$$D = \begin{vmatrix} +3 & -3 & +5 \\ +1 & +2 & -6 \\ +2 & -1 & +3 \end{vmatrix} = +20 \tag{5.20a}$$

$$D_1 = \begin{vmatrix} +4 & -3 & +5 \\ +3 & +2 & -6 \\ +1 & -1 & +3 \end{vmatrix} = +20 \tag{5.20b}$$

$$D_2 = \begin{vmatrix} +3 & +4 & +5 \\ +1 & +3 & -6 \\ +2 & +1 & +3 \end{vmatrix} = -40 \tag{5.20c}$$

$$D_3 = \begin{vmatrix} +3 & -3 & +4 \\ +1 & +2 & +3 \\ +2 & -1 & +1 \end{vmatrix} = -20 \tag{5.20d}$$

Then

$$x_1 = \frac{+20}{+20} = +1.00 \qquad x_2 = \frac{-40}{+20} = -2.00 \qquad x_3 = \frac{-20}{+20} = -1.00 \tag{5.21}$$

Before closing this discussion of Cramer's rule (or, as it is sometimes called, solution by determinants), it would be well to consider the "time of solution" implications of this method. A 5 × 5 set requires the carrying out of $6 \times 5! = 720$ multiplications. Counting just these multiplications and assuming that the computer to be used can carry out one multiplication in 300×10^{-6} sec shows that 0.216 sec would be required to solve the problem. While this may not seem to be too large a number, consider the rate at which time increases as the number of equations becomes larger, as shown in Table 5.1. It should be evident that this method is not used to solve larger problems.

Table 5.1

Order of equations	Time for multiplications			
	Seconds	Hours	Days	Years
5 × 5	0.216			
10 × 10	11.98×10^3	3.32	0.134	
15 × 15	6.277×10^9	1.745×10^6	72.7×10^3	199
20 × 20	5.110×10^{19}	4.26×10^{12}	17.8×10^{10}	4.85×10^8

IN 1969

NOW IN 1979 IT WIL
TAKE 2-3 seco

5.3 Elimination Methods

Methods involving the successive elimination of the unknowns constitute the major, direct way of solving simultaneous linear equations. There are many variations on this method, and a few of the more important of these will be presented.

5.3.1 Gauss' Method

STRONG DIAGONAL NOT NEE

To illustrate the basic features of Gauss' method, the most widely known of the elimination methods, consider the general 3 × 3 problem.

✓ OK REARRANGING

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= c_3 \end{aligned} \tag{5.22}$$

By assuming that $a_{11} \neq 0$, the first of these equations can be rewritten with a unit coefficient for x_1 by dividing the entire equation by a_{11}. That is,

②

$$x_1 + b_{12}x_2 + b_{13}x_3 = b_{14} \tag{5.23}$$

where

$$b_{12} = \frac{a_{12}}{a_{11}} \qquad b_{13} = \frac{a_{13}}{a_{11}} \qquad b_{14} = \frac{c_1}{a_{11}} \tag{5.24}$$

④

Multiplying Eq. (5.23) by a_{21} and then subtracting the resulting equation from the second of Eqs. (5.22) gives an expression in only x_2 and x_3:

$$a_{22-1}x_2 + a_{23-1}x_3 = c_{2-1} \tag{5.25}$$

The subscript notation 22 − 1 is used to indicate the original 2-2 position with a new set of coefficients in an $n - 1$ reduced system.

$$a_{21-1} = a_{22} - a_{21}b_{12} \qquad a_{23-1} = a_{23} - a_{21}b_{13} \qquad c_{2-1} = c_2 - a_{21}b_{14} \tag{5.26}$$

A similar multiplication of Eq. (5.23) by a_{31} and a subtraction from the third of Eqs. (5.22) yields

$$a_{32-1}x_2 + a_{33-1}x_3 = c_{3-1} \tag{5.27}$$

The coefficient expressions are given by

$$a_{32-1} = a_{32} - a_{31}b_{12} \qquad a_{33-1} = a_{33} - a_{31}b_{13} \qquad c_{3-1} = c_3 - a_{31}b_{14} \tag{5.28}$$

Equation (5.25) can now be divided by a_{22-1}:

$$x_2 + b_{23-1}x_3 = b_{24-1} \tag{5.29}$$

where $$b_{23-1} = \frac{a_{23-1}}{a_{22-1}} \qquad \text{and} \qquad b_{24-1} = \frac{c_{2-1}}{a_{22-1}} \tag{5.30}$$

Multiplying Eq. (5.29) by a_{32-1} and subtracting the resulting expression from Eq. (5.27) yields the following equation in the single variable x_3:

$$a_{33-2}x_3 = c_{3-2} \tag{5.31}$$

$$a_{33-2} = a_{33-1} - a_{32-1}b_{23-1} \qquad c_{3-2} = c_{3-1} - a_{32-1}b_{24-1} \tag{5.32}$$

Dividing Eq. (5.31) by a_{33-2},

$$x_3 = b_{34-2} \tag{5.33}$$

and $$b_{34-2} = \frac{c_{3-2}}{a_{33-2}} \tag{5.34}$$

Summarizing, the given three equations (5.22) have been rearranged in the triangular form

$$\begin{aligned} x_1 + b_{12}x_2 + b_{13}x_3 &= b_{14} \\ x_2 + b_{23-1}x_3 &= b_{24-1} \\ x_3 &= b_{34-2} \end{aligned} \tag{5.35}$$

It should be noted that, a value for x_3 having been found by the described forward-elimination procedure, values of x_2 and x_1 can be obtained by *back substitution.* Only one unknown will appear in each of the calculations.

An $n \times n$ system now will be discussed. It will be assumed that the equations have the form

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= c_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= c_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= c_n
\end{aligned}
\tag{5.36}
$$

For $a_{11} \neq 0$, the first of these equations can be written as

$$x_1 + b_{12}x_2 + b_{13}x_3 + \cdots + b_{1n}x_n = b_{1m} \tag{5.37}$$

where

$$b_{1j} = \frac{a_{1j}}{a_{11}} \qquad j > 1 \tag{5.38}$$

and

$$m = n + 1$$

COINCIDE
an NX
SYSTE
(SAME
PROCEE

An $(n - 1)$ set of equations in the unknowns $x_2, x_3, \ldots, x_n$ can be obtained by successive elimination between Eq. (5.37) and the latter $(n - 1)$ equations of (5.36). Dividing the coefficients of the first equation in this new set by the leading element a_{22-1} gives the following equation:

$$x_2 + b_{23-1}x_3 + \cdots + b_{2n-1}x_n = b_{2m-1} \tag{5.39}$$

where

$$b_{2j-1} = \frac{a_{2j-1}}{a_{22-1}} \qquad j > 2 \tag{5.40}$$

By continuing the process, the leading equation of the $(n - 2)$ system can be written as

$$x_3 + b_{34-2}x_4 + \cdots + b_{3n-2}x_n = b_{3m-2} \tag{5.41}$$

and

$$b_{3j-2} = \frac{a_{3j-2}}{a_{33-2}} \qquad j > 3 \tag{5.42}$$

Finally, there will have been carried out $(n - 1)$ eliminations. For this case

$$x_n = b_{nm-(n-1)} \tag{5.43}$$

where

$$b_{nm-(n-1)} = \frac{a_{nm-(n-1)}}{a_{nn-(n-1)}} \tag{5.44}$$

Back substitution will yield the other $(n - 1)$ x values.

Consider the solution by Gauss elimination of the numerical example solved in Sec. 5.2 by using Cramer's rule. The equations are repeated as (5.45).

$$
\begin{aligned}
3x_1 - 3x_2 + 5x_3 &= +4 \\
1x_1 + 2x_2 - 6x_3 &= +3 \\
2x_1 - 1x_2 + 3x_3 &= +1
\end{aligned}
\tag{5.45}
$$

The solution is shown in Table 5.2; a parallel part of the table indicates the coefficients described above.

Table 5.2

Line	x_1	x_2	x_3	= const	ΣH	x_1	x_2	x_3	= const	ΣH
1	a_{11}	a_{12}	a_{13}	c_1		+3	−3	+5	+4	+9.00
2	a_{21}	a_{22}	a_{23}	c_2		+1	+2	−6	+3	0
3	a_{31}	a_{32}	a_{33}	c_3		+2	−1	+3	+1	+5.00
4	1	b_{12}	b_{13}	b_{14}		1	−1.00	+1.67	+1.33	+3.00
5		a_{22-1}	a_{23-1}	c_{2-1}(or a_{24-1})			+3.00	−7.67	+1.67	−3.00
6		a_{32-1}	a_{33-1}	c_{3-1}(or a_{34-1})			+1.00	−0.34	−1.66	−1.00
7		1	b_{23-1}	b_{24-1}			1	−2.56	+0.56	−1.00
8			a_{33-2}	c_{3-2}(or a_{34-2})				+2.22	−2.22	0
9			1	$b_{34-2} = x_3$				1	−1.00	0
10		1					1		−2.00	
11	1					1			+1.00	

It should be noted that the arrangement of Table 5.2 affords a well-ordered system for carrying out the required calculations. To illustrate this, consider the first five lines of the left-hand part of Table 5.2, as shown in Table 5.3. The element a_{22-1} in line 5 is obtained by subtracting from a_{22} the product of a_{21} and b_{12}. The elements have been indicated by the

Table 5.3

Line	x_1	x_2	x_3	= const
1	a_{11}	a_{12}	a_{13}	c_1
2	$\boxed{a_{21}}$	a_{22}	a_{23}	$\boxed{c_2}$
3	a_{31}	a_{32}	a_{33}	c_3
4	1	b_{12}	b_{13}	$\boxed{b_{14}}$
5		a_{22-1}	a_{23-1}	$\boxed{c_{2-1}}$

dashed oval. The element c_{2-1} would be obtained in a similar manner by subtracting from c_2 the product of a_{21} and b_{14}.

The process just described is a continuing one. Line 8, for example, is determined from lines 5, 6, and 7, as shown in Table 5.4, which is a segment of the original table. The entry in the "Constant" column would be

$$(-1.66) - (+1.00)(+0.56) = -2.22 \tag{5.46}$$

Table 5.4

Line	x_1	x_2	x_3	= const	ΣH
5		+3.00	−7.67	+1.67	−3.00
6		$\boxed{+1.00}$	−0.34	$\boxed{-1.66}$	−1.00
7		1	−2.56	$\boxed{+0.56}$	−1.00
8			+2.22	$\boxed{-2.22}$	0

As is true in all tabular forms of progressive calculations, errors at an early stage are not readily apparent. Moreover, they may alter the solution materially. To afford checkpoints against such mistakes, it is desirable to add another column to the given set. This was done in Table 5.2, where the added column was labeled ΣH and is the sum of all elements to the left in the line in question. By dealing with this added column as if it were part of the original system, a check is made when the sum of all entries in the line equals that calculated by the above-defined process. For example, the ΣH entry in line 8 is obtained by subtracting from -1.00 (line 6) the product $(+1.00)(-1.00)$. This equals $(+2.22) + (-2.22) = 0$, and a check on the arithmetic is thus carried out. Had the two been different because of roundoff errors, it would be necessary to make a decision as to which of the two values is to be used. The average value is frequently the more desirable one. It should be recognized, however, that when any one of the lines does not add horizontally to that given as ΣH, subsequent entry will in all likelihood also not check.

It can be shown that a total of

$$\frac{n}{3}(n^2 + 3n - 1) \tag{5.47}$$

multiplications and divisions are required to solve a set of n linear simultaneous algebraic equations by the single-division Gauss elimination procedure described above. Since multiplication and division times on a computer are approximately equal, these can be lumped together when approximate times of solution are being determined. A 5 × 5 set requires 65 such operations. If 300×10^{-6} sec for one multiplication or division is assumed, as was done in the preceding section, the multiplication and division time required will be 0.0195 sec. The following table gives values for larger-order equations.

Order of equations	Time for multiplications and divisions, sec
5 × 5	0.020
10 × 10	0.129
15 × 15	0.404
20 × 20	0.918
40 × 40	6.876

5.3.2 Cholesky's Method

— ELIMINATION

A variation of the general elimination procedure that can be adapted readily to computer usage is that known as Cholesky's method, or Crout's method. In essence, it consists in the determination of a lower triangular matrix that will reduce the original system to the upper unit triangular form.

The original set of equations to be considered is

$$[A][X] = [C] \tag{5.48}$$

or

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{bmatrix} \tag{5.49}$$

If these can be redefined in the upper unit triangular form of Eqs. (5.50) or (5.51), the solution can be obtained by back substitution.

$$\begin{bmatrix} 1 & t_{12} & \cdots & t_{1n} \\ 0 & 1 & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} k_1 \\ k_2 \\ \cdot \\ \cdot \\ \cdot \\ k_n \end{bmatrix} \tag{5.50}$$

$$[T][X] = [K] \tag{5.51}$$

Since any nonsingular matrix (that is, a matrix having the value of the determinant of the coefficients other than zero) can be defined to be the product of two other nonsingular matrices, assume that the original $[A]$ is the result of a matrix multiplication of a lower triangular and an upper unit triangular matrix. That is,

$$[L][T] = [A] \tag{5.52}$$

This implies, however, that the constants are also related through $[L]$ according to Eq. (5.53):

$$[L][K] = [C] \tag{5.53}$$

This follows from the identity to zero of the two systems, as shown in Eq. (5.54):

$$[L][TX - K] = 0 = [AX - C] \tag{5.54}$$

Equation (5.54) suggests that for computational work the augmented form of the matrices may have advantages. By definition, the *augmented matrix* is obtained by adding to the square coefficient matrix the column vector of constants. The augmented matrix of Eq. (5.49) is given as Eq. (5.55):

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \cdots & \cdots & \cdots & \cdots & \cdot \\ & & & & \cdot \\ & & & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} & c_n \end{array}\right] \tag{5.55}$$

$$|\!\longleftarrow [A] \longrightarrow\!| \; [C] |\!\longleftarrow$$

By using this notation, Eqs. (5.52) and (5.53) can be written by one set of equations:

$$\underset{\substack{\text{Augmented matrix}\\ [A \mid C]}}{\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & c_1 \\ a_{21} & a_{22} & a_{23} & c_2 \\ a_{31} & a_{32} & a_{33} & c_3 \end{array}\right]} = \underset{[L]}{\begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}} \underset{\substack{\text{Augmented matrix}\\ [T \mid K]}}{\left[\begin{array}{ccc|c} 1 & t_{12} & t_{13} & k_1 \\ 0 & 1 & t_{23} & k_2 \\ 0 & 0 & 1 & k_3 \end{array}\right]} \tag{5.56}$$

Matrix multiplication can be used to define the entries ℓ_{ij}, t_{ij}, and k_i in terms of the original a's and c's. For example,

$$a_{11} = \ell_{11}(1) + (0)(0) + (0)(0) \qquad \text{or} \qquad \ell_{11} = a_{11} \tag{5.57}$$

The total group would be

1

$$\ell_{11} = a_{11} \qquad \ell_{21} = a_{21} \qquad \ell_{31} = a_{31}$$

4

$$\ell_{22} = a_{22} - \ell_{21}t_{12} \qquad \ell_{32} = a_{32} - \ell_{31}t_{12}$$

7

$$\ell_{33} = a_{33} - \ell_{31}t_{13} - \ell_{32}t_{23}$$

2

$$t_{12} = \frac{a_{12}}{\ell_{11}} \qquad t_{13} = \frac{a_{13}}{\ell_{11}}$$

5

$$t_{23} = \frac{a_{23} - \ell_{21}t_{13}}{\ell_{22}}$$

3

$$k_1 = \frac{c_1}{\ell_{11}}$$

6

$$k_2 = \frac{c_2 - \ell_{21}k_1}{\ell_{22}}$$

8

$$k_3 = \frac{c_3 - \ell_{31}k_1 - \ell_{32}k_2}{\ell_{33}} \tag{5.58}$$

It should be noted that if the following order of computation is carried out, all entries will be defined in terms of previously calculated values.

1. Calculate first column of ℓ's; calculate first row of t's and k_1.
2. Calculate second column of ℓ's; calculate second row of t's and k_2.
3. Calculate ℓ_{33}; calculate k_3.

The unknown x's are then obtained by back substitution [using Eqs. (5.50)].

$$\begin{aligned} x_3 &= k_3 \\ x_2 &= k_2 - t_{23}x_3 \\ x_1 &= k_1 - t_{13}x_3 - t_{12}x_2 \end{aligned} \tag{5.59}$$

Since this method assumes an upper unit triangular matrix in all cases, space may be saved and a computational advantage realized by eliminating from the tabulation the 1's from $[T]$ and writing $[T]$ and $[L]$ in a square-set form. However, to emphasize that these are in fact two separate matrices, it is suggested that a heavy "stair-type" separation be shown. Equations (5.56) would be listed as shown in Eqs. (5.60).

$$\left[\begin{array}{ccc:c} a_{11} & a_{12} & a_{13} & c_1 \\ a_{21} & a_{22} & a_{23} & c_2 \\ a_{31} & a_{32} & a_{33} & c_3 \end{array}\right] = \left[\begin{array}{c|cc:c} \ell_{11} & t_{12} & t_{13} & k_1 \\ \cline{2-3} \ell_{21} & \ell_{22} & \multicolumn{1}{|c:}{t_{23}} & k_2 \\ \cline{3-3} \ell_{31} & \ell_{32} & \ell_{33} & \multicolumn{1}{|c}{k_3} \end{array}\right] \tag{5.60}$$

These can also be shown in the form of Table 5.5. To illustrate the computational advantage of using Table 5.5, consider the determination of

Table 5.5

$i \backslash j$	1 (x_1)	2 (x_2)	3 (x_3)	4 (Const)	ΣH	1	2	3	4	ΣH
1	a_{11}	a_{12}	a_{13}	c_1		ℓ_{11}	t_{12}	t_{13}	k_1	
2	a_{21}	a_{22}	a_{23}	c_2		ℓ_{21}	ℓ_{22}	t_{23}	k_2	
3	a_{31}	a_{32}	a_{33}	c_3		ℓ_{31}	ℓ_{32}	ℓ_{33}	k_3	

$[A \mid C]$ $[L \backslash T \mid K]$

ℓ_{22}. Reference to Eqs. (5.58) shows that

$$\ell_{22} = a_{22} - \ell_{21}t_{12}$$

That is, ℓ_{22} is found by subtracting from a_{22} the product of the leading row and column elements defining the 2-2 position in the LT set. The element ℓ_{33} equals a_{33} minus the sum of the paired products $\ell_{31}t_{13}$ and $\ell_{32}t_{23}$. Element t_{23}, on the other hand, is obtained by first subtracting from a_{23} the quantity $\ell_{21}t_{13}$ and then dividing the remainder by ℓ_{22}. It should be noted that operations of this type can be carried out on most modern desk calculators without reentering intermediate values.

The 3 × 3 systems having been discussed and an ordered calculational procedure having been defined, extension to the $n \times n$ case can be readily made. In general terms

$$\ell_{ij} = a_{ij} - \sum_{k=1}^{j-1} \ell_{ik}t_{kj} \qquad \ell_{i1} = a_{i1} \qquad i > j \tag{5.61}$$

$$t_{ij} = \frac{1}{\ell_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} \ell_{ik}t_{kj}\right) \qquad t_{1j} = \frac{a_{1j}}{a_{11}} \qquad i < j \tag{5.62}$$

where the order in which calculations are carried out is by rows (for a computer) or as shown in the following table (for a desk calculator):

ℓ_{11}	t_{12}	t_{13}	$\cdots$	t_{1n}	k_1	first set of calculations
ℓ_{21}	ℓ_{22}	t_{23}	$\cdots$	t_{2n}	k_2	second set of calculations
ℓ_{31}	ℓ_{32}	ℓ_{33}	$\cdots$	t_{3n}	k_3	third set of calculations
$\cdot$	$\cdot$	$\cdot$		$\cdot$	$\cdot$	
$\cdot$	$\cdot$	$\cdot$		$\cdot$	$\cdot$	
$\cdot$	$\cdot$	$\cdot$		$\cdot$	$\cdot$	
ℓ_{n1}	ℓ_{n2}	ℓ_{n3}	$\cdots$	ℓ_{nn}	k_n	

Again consider for illustration the three simultaneous linear algebraic equations that were earlier solved by Cramer's rule and Gauss elimination, that is,

$$\begin{aligned} 3x_1 - 3x_2 + 5x_3 &= +4 \\ 1x_1 + 2x_2 - 6x_3 &= +3 \\ 2x_1 - 1x_2 + 3x_3 &= +1 \end{aligned} \tag{5.63}$$

A tabular solution is given in Table 5.6. Back substitution yields the following values for the unknowns:

Table 5.6

i \ j	1	2	3	4	ΣH	1	2	3	4	ΣH†
1	+3.00	−3.00	+5.00	+4.00	+9.00	+3.00	−1.00	+1.67	+1.33	+3.00
2	+1.00	+2.00	−6.00	+3.00	0	+1.00	+3.00	−2.56	+0.56	−1.00
3	+2.00	−1.00	+3.00	+1.00	+5.00	+2.00	+1.00	+2.23	−1.00	0

† It should be noted that this horizontal summation is only for the $[T \vdots K]$ augmented matrix. Elements of $[L]$ should not be included. Also, the implied 1:00 in the ii position of the $[T]$ matrix must be added when summing horizontally to allow a check.

$$\begin{aligned} x_3 &= -1.00 \\ x_2 &= +0.56 - (-2.56)(x_3) = -2.00 \\ x_1 &= +1.33 - (+1.67)(x_3) - (-1.00)(x_2) = +1.00 \end{aligned} \tag{5.64}$$

Again consider the total number of multiplications and divisions required to obtain a solution by using Crout's method. The number of these operations required at each stage for a 5 × 5 set is shown in the following Table 5.7.

Table 5.7

Line	1	2	3	4	5	6
1	0	1	1	1	1	1
2	0	1	2	2	2	2
3	0	1	2	3	3	3
4	0	1	2	3	4	4
5	0	1	2	3	4	5

Back substitution will require an additional

$$0 + 1 + 2 + 3 + 4 = 10 \tag{5.65}$$

The total number is then 55 (from Table 5.7) plus 10 from back substitution, or a total of 65 multiplications and divisions. This is exactly the same number as was found for Gauss elimination. A comparison of the equations defining the entries in each system indicates that the number of multiplications and divisions should be equal, the identical operations having been carried out. The advantage of Crout's method over Gauss' is then the compactness of form and the elimination of entry of intermediate steps. This can be seen from Tables 5.2 and 5.6.

It should be understood that these time calculations are only for multiplications and divisions. While, in general, these require much more computational time than additions and subtractions, should the number of additions and subtractions become large in comparison with the multiplications and divisions, a realistic time evaluation must include both items.

5.3.3 Square-root Method

WE CAN USE CHOLESKY'S METHO
INSTEAD IF W
WAN

When the coefficient matrix is symmetrical, the method due to Banachiewicz, known as the square-root method, is one of the most efficient means of solving simultaneous linear algebraic equations. Here, as in Crout's method, the augmented matrix $[A \vdots C]$ is assumed to be the product of two other matrices. These latter calculational matrices, however, are presupposed to be an upper triangular matrix $[U]$ and its transpose $[U']$,

$$[A \vdots C] = [U'][U \vdots K] \tag{5.66}$$

or, in expanded form,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ u_{12} & u_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ u_{1n} & u_{2n} & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \tag{5.67}$$

and

$$\begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{bmatrix} = \begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ u_{12} & u_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ u_{1n} & u_{2n} & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \cdot \\ \cdot \\ \cdot \\ k_n \end{bmatrix} \tag{5.68}$$

From matrix multiplication

$$\begin{aligned} a_{ij} &= u_{1i}u_{1j} + u_{2i}u_{2j} + \cdots + u_{ii}u_{ij} && i < j \\ a_{ii} &= u_{1i}^2 + u_{2i}^2 + \cdots + u_{ii}^2 && i = j \end{aligned} \tag{5.69}$$

Similar equations relate the u's, k's, and c's.

By solving Eqs. (5.69) for the values of u_{ij} in terms of the given a's, Eqs. (5.70) to (5.72) are obtained.

$$u_{11} = \sqrt{a_{11}} \qquad u_{1j} = \frac{a_{1j}}{u_{11}} \tag{5.70}$$

$$u_{ii} = \sqrt{a_{ii} - \sum_{p=1}^{i-1} u_{pi}^2} \quad i > 1 \qquad u_{ij} = \frac{1}{u_{ii}}\left(a_{ij} - \sum_{p=1}^{i-1} u_{pi}u_{pj}\right) \quad j > 1 \tag{5.71}$$

and

$$u_{ij} = 0 \qquad i > j \tag{5.72}$$

The u's having been defined, the k's are now readily determined.

$$k_1 = \frac{c_1}{u_{11}} \qquad k_i = \frac{1}{u_{ii}}\left(c_i - \sum_{p=1}^{i-1} u_{pi}k_p\right) \qquad i > 1 \tag{5.73}$$

Back substitution gives for values of the unknown x's

$$x_n = \frac{k_n}{u_{nn}} \qquad \text{and} \qquad x_i = \frac{k_i - \sum_{p=i+1}^{n} u_{ip}x_p}{u_{ii}} \qquad i < n \tag{5.74}$$

A tabular arrangement of the type shown in Table 5.8 can be used to carry out calculations by the square-root method. Because of assumed symmetry, only half the original $[A]$ is given. Similarly, only $[U]$ is shown. The u and k values for the illustrated 3×3 system are those given as Eqs. (5.75a) to (5.75c).

Table 5.8

j / i	1 (x_1)	2 (x_2)	3 (x_3)	4 (Const)	ΣH	1	2	3	4	ΣH
1	a_{11}	a_{12}	a_{13}	c_1		u_{11}	u_{12}	u_{13}	k_1	
2		a_{22}	a_{23}	c_2			u_{22}	u_{23}	k_2	
3			a_{33}	c_3				u_{33}	k_3	

$$u_{11} = \sqrt{a_{11}} \qquad u_{12} = \frac{a_{12}}{u_{11}} \qquad u_{13} = \frac{a_{13}}{u_{11}} \qquad k_1 = \frac{c_1}{u_{11}} \tag{5.75a}$$

$$u_{22} = \sqrt{a_{22} - u_{12}^2} \qquad u_{23} = \frac{a_{23} - u_{12}u_{13}}{u_{22}} \qquad k_2 = \frac{c_2 - u_{12}k_1}{u_{22}} \tag{5.75b}$$

$$u_{33} = \sqrt{a_{33} - (u_{13}^2 + u_{23}^2)} \qquad k_3 = \frac{c_3 - (u_{13}k_1 + u_{23}k_2)}{u_{33}} \tag{5.75c}$$

The unknown x's are

$$x_3 = \frac{k_3}{u_{33}} \qquad x_2 = \frac{k_2 - u_{23}x_3}{u_{22}} \qquad x_1 = \frac{k_1 - (u_{12}x_2 + u_{13}x_3)}{u_{11}} \cdot \tag{5.76}$$

As a numerical example, consider the following symmetrical, tridiagonal set of algebraic equations:

$$\begin{bmatrix} +4 & +2 & 0 \\ +2 & +10 & +4 \\ 0 & +4 & +5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +2 \\ +6 \\ +5 \end{bmatrix} \tag{5.77}$$

The tabular solution is given in Table 5.9.

Table 5.9

SKIP

i \ j	1 (x_1)	2 (x_2)	3 (x_3)	4 (const)	ΣH†	1	2	3	4	ΣH
1	+4	+2	0	+2	+8	+2.00	+1.00	0	+1.00	+4.00
2		+10	+4	+6	+22		+3.00	+1.33	+1.67	+6.00
3			+5	+5	+14			+1.80	+1.54	+3.34

† Must include in ΣH the $a_{ij} = a_{ji}$ terms.

$$\begin{aligned} x_3 &= \frac{+1.54}{+1.80} = +0.855 \\ x_2 &= \frac{+1.67 - (+1.33)(+0.855)}{+3.00} = +0.177 \\ x_1 &= \frac{+1.00 - (+1.00)(+0.177) - (0)(+0.855)}{+2.00} = +0.412 \end{aligned} \tag{5.78}$$

The student is urged to verify each of the entries in Table 5.9.

5.4 Iteration Methods

STRONG DIAGONALS
ONLY

The methods for solving simultaneous linear algebraic equations that have been discussed thus far require the complete set of calculations to be carried out before even an approximate answer to the group can be found. It should also be apparent that these methods are highly sensitive to computational errors. Another feature that may be considered objectionable is that little if any computational advantage can be realized for zeros that may be contained in the given set when elimination methods are used.

One means of obtaining an approximate solution to a given system is to make an "educated guess." If the first assumed values do not work, select others. By carefully noting the influence of each variable, these selections need not be haphazard, and the correct set of values for the system of equations can be approached. It should be recognized that when this procedure is used, an error in any one guess does not prevent the correct solution from being obtained even though a few more trials may be required to obtain the necessary accuracy.

When guessing follows a prescribed pattern—the assumed values for the $k + 1$ trial being the resulting values of trial k—the method of obtaining a solution is known as *iteration*. Two iterative methods that will be discussed are the Jacobi method and the Gauss-Seidel method.

For iteration to be successful, the coefficient matrix of the given system of equations must be such that successive approximations converge to the correct solution. In general, *a necessary condition for convergence is that the set be diagonal.* This requires that one of the coefficients in each of the equations be greater than all others and that this "strong coefficient" be contained in a different position in each equation of the set. (Under these conditions, the system of equations could, if desired, be rearranged with all strong elements in the ii position.) A second condition that is *sufficient to ensure convergence* exists when the absolute value of the strong element in each of the equations is greater than the sum of the absolute values of all other coefficient elements in that line. There are other less stringent requirements for convergence; however, these are beyond the scope of this introductory work.

FFICIENT
QUIREMENTS

5.4.1 Jacobi's Method — ITERATIVE METHOD —

A strong diagonal set of linear algebraic equations, as represented by Eqs. (5.79), can be written in the matrix form of Eqs. (5.80), where the unknown associated with the strong diagonal element has been isolated for solution.

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= c_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= c_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= c_n
\end{aligned}
\tag{5.79}
$$

$$
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}
=
\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \cdot \\ \cdot \\ \cdot \\ d_n \end{bmatrix}
-
\begin{bmatrix}
0 & b_{12} & b_{13} & \cdots & b_{1n} \\
b_{21} & 0 & b_{23} & \cdots & b_{2n} \\
b_{31} & b_{32} & 0 & \cdots & b_{3n} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
 & & & & \\
b_{n1} & b_{n2} & b_{n3} & \cdots & 0
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}
\tag{5.80}
$$

$$
d_i = \frac{c_i}{a_{ii}} \qquad \text{and} \qquad b_{ij} = \frac{a_{ij}}{a_{ii}} \tag{5.81}
$$

At first glance it would seem that Eqs. (5.80) are inconsistent, x_i appearing on both sides of the equation. If it is noted that zeros are contained in all ii positions in $[B]$, there is correspondence with the original set. A more meaningful statement of Eqs. (5.80) might be

$$_{(k+1)}[X] = [D] - [B]_{(k)}[X] \tag{5.82}$$

where k represents the number of the iteration and $_{(k)}[X]$ denotes the x values obtained after k trials. $_{(k+1)}[X]$ represents the computed (new) values of the unknown x's.

Consider for illustration the numerical example of Sec. 5.3.3. The equations are repeated as (5.83).

STRONG
DIAGONAL

$$\begin{aligned} \underline{4x_1} + 2x_2 + 0x_3 &= +2 \\ 2x_1 + \underline{10x_2} + 4x_3 &= +6 \\ 0x_1 + 4x_2 + \underline{5x_3} &= +5 \end{aligned} \tag{5.83}$$

In the iteration normal form of Eqs. (5.80), these would be written as

$$\begin{array}{ccccccc} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} & = & \begin{bmatrix} +0.50 \\ +0.60 \\ +1.00 \end{bmatrix} & - & \begin{bmatrix} 0 & +0.50 & 0 \\ +0.20 & 0 & +0.40 \\ 0 & +0.80 & 0 \end{bmatrix} & \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ _{(k+1)}[X] & = & [D] & - & [B] & _{(k)}[X] \end{array} \tag{5.84}$$

It should be noted that the set of equations being considered has a strong diagonal. Moreover, the sum of the absolute values of all elements in each row of $[B]$ is less than 1.0. Iteration therefore will converge to the correct solution.

Assume, as a starting point, that $x_1 = x_2 = x_3 = 0$. (Any other values could just as well have been selected.) From Eqs. (5.84)

$$x_1 = +0.50 \qquad x_2 = +0.60 \qquad x_3 = +1.00 \tag{5.85}$$

Now, by using these computed values as a guess, a new set of x's can be obtained from Eqs. (5.84).

$$\begin{aligned} x_1 &= +0.50 - (0.50)(+0.60) = +0.20 \\ x_2 &= +0.60 - (0.20)(+0.50) - (0.40)(+1.00) = +0.10 \\ x_3 &= +1.00 - (0.80)(+0.60) = +0.52 \end{aligned} \tag{5.86}$$

Table 5.10 is a listing of the values obtained for the three x's at different stages of iteration. A plot of these values is given in Fig. 5.2.

5.4.2 Gauss-Seidel Method

— ITERATION METHOD

To afford a more rapid convergence of the iterative process, the Gauss-Seidel method presupposes the most recent cycle (or part of a cycle) of

Table 5.10

Cycle:	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth
x_1	+0.50	+0.20	+0.450	+0.324	+0.429	+0.376	+0.420	+0.398	+0.417	+0.412
x_2	+0.60	+0.10	+0.352	+0.142	+0.248	+0.160	+0.204	+0.167	+0.186	+0.170
x_3	+1.00	+0.52	+0.920	+0.718	+0.886	+0.802	+0.872	+0.837	+0.866	+0.851

values at each stage of calculations. For example, assume for a 5 × 5 set that a value for x_3 is being sought for the fourth iteration. Values of x_4 and x_5 to be used in the calculations are those associated with the third iteration. Fourth iterative values of x_1 and x_2 would be used. Again for illustration, consider the set of equations (5.83) [or (5.84), which is in the normal form for iteration]. Starting with assumed values of $x_2 = x_3 = 0$,

$$x_1 = +0.50 \tag{5.87}$$

By using this new value for x_1 and keeping $x_3 = 0$, x_2 would be

$$x_2 = +0.60 - (0.20)(+0.50) = +0.50 \tag{5.88}$$

In turn,

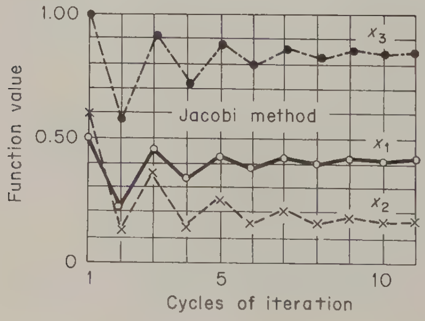
$$x_3 = +1.00 - (0.80)(+0.50) = +0.60 \tag{5.89}$$

x_1 is now recomputed from the first of Eqs. (5.84), assuming $x_2 = +0.50$ and $x_3 = +0.60$. The solution is summarized in Table 5.11, and the values are plotted in Fig. 5.3.

Table 5.11

Cycle:	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth
x_1	+0.50	+0.25	+0.345	+0.384	+0.401	+0.408	+0.411	+0.412	+0.413
x_2	+0.50	+0.31	+0.231	+0.197	+0.183	+0.177	+0.175	+0.174	+0.173
x_3	+0.60	+0.75	+0.815	+0.842	+0.854	+0.858	+0.858	+0.861	+0.862

Figure 5.2



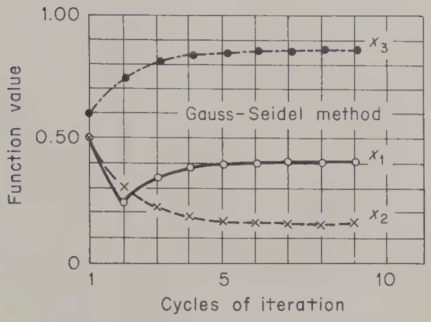


Figure 5.3

5.4.3 Iteration of Nondiagonal Sets

Before leaving the discussion of iteration methods it would be desirable to observe the behavior of a set that is *not diagonal.* Consider the equations that were solved earlier in this chapter by Cramer's rule and by Gauss elimination.

$$\begin{aligned} \underline{3x_1} \quad -3x_2 \quad +5x_3 &= +4 \\ 1x_1 \quad +\underline{2x_2} \quad -6x_3 &= +3 \\ 2x_1 \quad -1x_2 \quad +\underline{3x_3} &= +1 \end{aligned} \tag{5.90}$$

In the normal form for iteration, these are written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +1.33 \\ +1.50 \\ +0.33 \end{bmatrix} - \begin{bmatrix} 0 & -1.00 & +1.67 \\ +0.50 & 0 & -3.00 \\ +0.67 & -0.33 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{5.91}$$

Using the Jacobi method, the first nine iterations are summarized in Table 5.12 and plotted in Fig. 5.4. The correct solutions are $x_1 = +1.00$, $x_2 = -2.00$, and $x_3 = -1.00$. It should be recalled that iteration is not recommended for these types of problems.

Table 5.12

Cycle:	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth
x_1	+1.33	+2.283	+3.425	+2.057	+2.765	+1.474	+2.023	−6.025	+6.909
x_2	+1.50	+1.825	−0.243	−1.953	−5.622	−4.950	−7.915	−0.514	−6.450
x_3	+0.33	−0.160	−0.580	−2.031	−1.689	−3.384	−0.334	−3.654	+4.179

5.5 Relaxation Methods

The method of relaxation (like iteration) differs from elimination in that solutions for all unknowns are obtained simultaneously. However, unlike

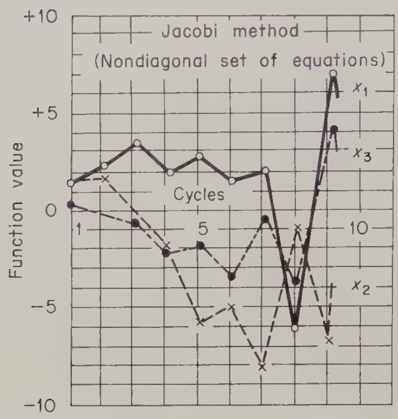


Figure 5.4

iteration, it is impossible in relaxation to prescribe at the outset a position-fixed order by which the various terms in the solution are to be "adjusted." Another variation from the iterative procedure is that relaxation usually is concerned with manipulation of the amount of "out-of-balance," rather than with the entire magnitude of the quantities in question at each stage.

The set of linear simultaneous algebraic equations to be considered is defined by Eq. (5.92).

$$[A][X] = [C] \tag{5.92}$$

These equations could have been written equally well as

$$[A][X] - [C] = 0 \tag{5.93}$$

or in the summation form

$$\sum_{j=1}^{n} a_{ij}x_j - c_i = 0 \qquad (i = 1, 2, \ldots, n) \tag{5.94}$$

For the correct solution of the x's, the equations will be satisfied identically. Deviations from these correct values, however, result in other than a "zero" correspondence. To facilitate a solution by the method of relaxation, Eqs. (5.94) will be written as

$$[A][X] - [C] = [R] \tag{5.95}$$

or

$$\sum_{j=1}^{n} a_{ij}x_j - c_i = R_i \qquad (i = 1, 2, \ldots, n) \tag{5.96}$$

or

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n - c_1 = R_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n - c_2 = R_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n - c_n = R_n \end{array} \tag{5.97}$$

The object is to alter the assumed x's so that the R's, *the residuals*, approach zero.

The residuals, as shown by Eqs. (5.95) to (5.97), are functions of the variable x's. That is,

$$R_i = R_i(x_1,x_2, \ldots ,x_n) \tag{5.98}$$

A change in the value of any one or all of the x's will alter the R's. By considering only the first term of the series describing the change in R with x, the total change in the residuals (that is, the differentials of the R's) will be those given by Eqs. (5.99).

$$\begin{aligned}
dR_1 &= \frac{\partial R_1}{\partial x_1}\,dx_1 + \frac{\partial R_1}{\partial x_2}\,dx_2 + \cdots + \frac{\partial R_1}{\partial x_n}\,dx_n \\
dR_2 &= \frac{\partial R_2}{\partial x_1}\,dx_1 + \frac{\partial R_2}{\partial x_2}\,dx_2 + \cdots + \frac{\partial R_2}{\partial x_n}\,dx_n \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
dR_n &= \frac{\partial R_n}{\partial x_1}\,dx_1 + \frac{\partial R_n}{\partial x_2}\,dx_2 + \cdots + \frac{\partial R_n}{\partial x_n}\,dx_n
\end{aligned} \tag{5.99}$$

However, from Eqs. (5.96) the partials can be determined:

$$\frac{\partial R_i}{\partial x_j} = a_{ij} \qquad (i, j = 1, 2, \ldots , n) \tag{5.100}$$

Equations (5.99) can therefore be written as

$$\begin{aligned}
dR_1 &= a_{11}\,dx_1 + a_{12}\,dx_2 + \cdots + a_{1n}\,dx_n \\
dR_2 &= a_{21}\,dx_1 + a_{22}\,dx_2 + \cdots + a_{2n}\,dx_n \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
dR_n &= a_{n1}\,dx_1 + a_{n2}\,dx_2 + \cdots + a_{nn}\,dx_n
\end{aligned} \tag{5.101}$$

The relative influence of each of the dx_i's on the residuals R_i now can be obtained by selecting, in turn, a unit change in the value of one x while all the others are kept unchanged. The corresponding unit-relaxation table can be set up as shown in Table 5.13, where in order to represent the finite increments involved, deltas replace corresponding total differentials in the R's. It should be noted that the entries in the ΔR portion of Table 5.13 are the transpose of the original a_{ij} coefficient matrix.

5.5.1 Relaxation of 3 × 3 Set of Equations

To illustrate how relaxation is carried out by using Table 5.13, consider the following set of equations. These are the same equations that were considered in Sec. 5.4, Iteration Methods.

Table 5.13 Unit-relaxation Table

Δx_1	Δx_2	$\cdots$	Δx_n	ΔR_1	ΔR_2	ΔR_3	$\cdots$	ΔR_n
1	0	$\cdots$	0	a_{11}	a_{21}	a_{31}	$\cdots$	a_{n1}
0	1	$\cdots$	0	a_{12}	a_{22}	a_{32}	$\cdots$	a_{n2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
0	0	$\cdots$	1	a_{1n}	a_{2n}	a_{3n}	$\cdots$	a_{nn}

$$\begin{aligned} 4x_1 \quad +2x_2 \qquad\qquad &= +2 \\ 2x_1 \quad +10x_2 \quad +4x_3 &= +6 \\ 4x_2 \quad +5x_3 &= +5 \end{aligned} \tag{5.102}$$

The corresponding unit-relaxation table is shown as Table 5.14. Since the given set is symmetric, the ΔR values in this table are also symmetric.

Table 5.14 Unit-relaxation Table

Δx_1	Δx_2	Δx_3	ΔR_1	ΔR_2	ΔR_3
1	0	0	$\underline{+4}$	+2	0
0	1	0	+2	$\underline{+10}$	+4
0	0	1	0	+4	$\underline{+5}$

TRANSPOSE

The standard form of the equations for relaxation is given by Eqs. (5.97). For the set in question, this is

$$\begin{aligned} 4x_1 \quad +2x_2 \qquad\qquad -2 &= R_1 \\ 2x_1 \quad +10x_2 \quad +4x_3 \quad -6 &= R_2 \\ 4x_2 \quad +5x_3 \quad -5 &= R_3 \end{aligned} \tag{5.103}$$

As a starting point, select $x_1 = x_2 = x_3 = 0$. (Any other values could have been presumed just as well.) The residuals are $R_1 = -2$, $R_2 = -6$, and $R_3 = -5$. Since the object of the relaxation procedure is to adjust the x's successively so that the R's approach zero, and since R_2 is the residual of greatest absolute magnitude, it is necessary first to determine which variable has the greatest influence on R_2 and then to determine how much that variable should be adjusted. From Table 5.14, it is observed that a unit change in x_2 will have the greatest influence on R_2, altering the value of R_2 by a factor of 10. Therefore a change of $\boxed{0.6}$ from the assumed value of $x_2 = 0$ will bring the residual R_2 to zero. At the same time, it will change the value of the residuals R_1 and R_3.

$\Delta R_2 \leftarrow$ COLUMN VE

$$\begin{aligned} R_1 &= -2 + \boxed{(2)}(0.6) = -0.8 \\ R_2 &= \cancel{0} -6 + \boxed{10}(0.6) = 0 \\ R_3 &= -5 + \boxed{(4)}(0.6) = -2.6 \end{aligned} \tag{5.104}$$

A second correction is now in order. Since R_3 is the greatest residual and since a variation in x_3 will have the greatest influence on R_3, a change in x_3 from zero to a value of $+0.52$ will reduce R_3 to zero. The residuals are now

$$\begin{aligned} R_1 &= -0.8 + \boxed{0}(0.52) = -0.8 \\ R_2 &= 0 + \boxed{(4)}(0.52) = +2.08 \\ R_3 &= -2.6 + \boxed{(5)}(0.52) = 0 \end{aligned} \tag{5.105}$$

At this stage, then, the approximate solutions with corresponding residuals are

$$\begin{aligned} x_1 &= 0 & &= 0 & R_1 &= -0.8 \\ x_2 &= 0 & +0.6 &= +0.6 & R_2 &= +2.08 \\ x_3 &= 0 & +0.52 &= +0.52 & R_3 &= 0 \end{aligned} \tag{5.106}$$

Table 5.15 summarizes the solution of the set of equations using this method. In all cases, the residual being reduced is underscored.

Table 5.15

Line	$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$R_1 = -2$	$R_2 = \underline{-6}$	$R_3 = -5$
1	0	+0.6	0	−0.8	0	$\underline{-2.6}$
2	0	0	+0.52	−0.8	$\underline{+2.08}$	0
3	0	−0.21	0	$\underline{-1.22}$	−0.02	−0.84
4	+0.31	0	0	+0.02	+0.60	$\underline{-0.84}$
5	0	0	+0.17	+0.02	$\underline{+1.28}$	+0.01
6	0	−0.13	0	−0.24	−0.02	$\underline{-0.51}$
7	0	0	+0.10	−0.24	$\underline{+0.38}$	−0.01
8	0	−0.04	0	$\underline{-0.32}$	−0.02	−0.17
9	+0.08	0	0	0	+0.14	$\underline{-0.17}$
10	0	0	+0.04	0	$\underline{+0.30}$	+0.03
11	0	−0.03	0	−0.06	0	$\underline{-0.09}$
12	0	0	+0.02	−0.06	$\underline{+0.08}$	+0.01
13	0	−0.01	0	$\underline{-0.08}$	−0.02	−0.03
14	+0.02	0	0	0	+0.02	$\underline{-0.03}$
15	0	0	+0.01	0	$\underline{+0.06}$	+0.02
16	0	−0.01	0	−0.02	−0.04	−0.02
Solution	+0.41	+0.17	+0.86	−0.02	−0.04	−0.02

To check that no arithmetic mistakes have been made, substitution of these values in Eqs. (5.103) yields

$$\begin{aligned}(4)(+0.41) + (2)(+0.17) - 2 &= R_1 = -0.02\\ (2)(+0.41) + (10)(+0.17) + (4)(+0.86) - 6 &= R_2 = -0.04\\ (4)(+0.17) + (5)(+0.86) - 5 &= R_3 = -0.02\end{aligned} \tag{5.107}$$

A consideration of Table 5.14, the unit-relaxation table for the problem in question, indicates that a unit variation in x_1 has its greatest influence on R_1, only half that value on R_2, and no influence whatsoever on R_3. The variable x_2 manifests itself most forcibly on R_2, R_1 and R_3 being much less affected. However, a unit change in x_3 has about the same influence on R_2 as it does on R_3. This latter situation is not too desirable for efficient relaxation. By adding, subtracting, and/or multiplying the individual rows of that unit-relaxation table, this situation can be improved and a more efficient operation table developed. A possible resulting *group-relaxation operation table* for the problem just considered is shown in Table 5.16. The entries in this table are obtained by multiplying row 3 of Table 5.14 by 2 and then subtracting row 2 of the same table from it. By using this group table, the solution is that shown in Table 5.17. Three relaxations using the group table 5.16 result in the same order of accuracy as did 16 relaxations using the unit-relaxation Table 5.14.

Table 5.16

Δx_1	Δx_2	Δx_3	ΔR_1	ΔR_2	ΔR_3
+1	0	0	$\underline{+4}$	+2	0
0	+1	0	+2	$\underline{+10}$	+4
0	−1	+2	−2	−2	$\underline{+6}$

Table 5.17

Line	$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$R_1 = -2$	$R_2 = \underline{-6}$	$R_3 = -5$
1	0	+0.6	0	−0.8	0	$\underline{-2.6}$
2	0	−0.43	+0.86	$\underline{-1.66}$	−0.86	−0.02
3	+0.42	0	0	+0.02	−0.02	−0.02
Solution	+0.42	+0.17	+0.86	+0.02	−0.02	−0.02

5.5.2 Relaxation of Larger Sets of Equations

A more realistic type of engineering problem might require the solution of the following set of equations:

$$
\begin{array}{lllll}
1.6x_1 & +2.0x_2 & +0.8x_3 & -1.2x_4 & = +37.2 \\
1.8x_1 & +0.7x_2 & -1.6x_3 & -0.8x_4 & = +42.0 \\
2.0x_1 & -1.6x_2 & +0.4x_3 & -3.0x_4 & = +13.8 \\
1.2x_1 & -1.4x_2 & +1.4x_3 & +2.0x_4 & = -30.4
\end{array}
\tag{5.108}
$$

Since any or all of these equations could be premultiplied throughout without changing their values (that is, they could be *scaled*), the equations will be written in the relaxation form:

$$
\begin{array}{llllll}
16x_1 & +20x_2 & +8x_3 & -12x_4 & -372 & = R_1 \\
18x_1 & +7x_2 & -16x_3 & -8x_4 & -420 & = R_2 \\
20x_1 & -16x_2 & +4x_3 & -30x_4 & -138 & = R_3 \\
12x_1 & -14x_2 & +14x_3 & +20x_4 & +304 & = R_4
\end{array}
\tag{5.109}
$$

The unit-relaxation table is given as Table 5.18.

Table 5.18 Unit-relaxation Table

Δx_1	Δx_2	Δx_3	Δx_4	ΔR_1	ΔR_2	ΔR_3	ΔR_4
1	0	0	0	+16	+18	+20	+12
0	1	0	0	+20	+7	−16	−14
0	0	1	0	+8	−16	+4	+14
0	0	0	1	−12	−8	−30	+20

It should be noted that none of the elements are particularly strong in this relaxation table. All values are approximately equal and, should relaxation be undertaken using them, convergence would be relatively slow. The development of a group-relaxation table is, therefore, highly desirable. If only unit combinations of the above are considered, Table

Table 5.19 Group-relaxation Table (1)

Δx_1	Δx_2	Δx_3	Δx_4	ΔR_1	ΔR_2	ΔR_3	ΔR_4
+1	+1	+1	0	$\underline{+44}$	+9	+8	+12
+1	0	−1	0	+8	$\underline{+34}$	+16	−2
0	+1	0	+1	+8	−1	$\underline{-46}$	+6
+1	0	0	+1	+4	+10	−10	$\underline{+32}$

5.19 is a possible group table. A "stronger" group table, however, is given in Table 5.20. It will be used in the solution that follows.

Table 5.20 Group-relaxation Table (2)

Δx_1	Δx_2	Δx_3	Δx_4	ΔR_1	ΔR_2	ΔR_3	ΔR_4
+1	+2	+2	0	$\underline{+72}$	0	−4	+12
−2	0	+3	−1	+4	$\underline{-76}$	+2	−2
0	+3	−1	+3	+16	+13	$\underline{-142}$	+4
+3	−1	+3	+3	+16	−25	−2	$\underline{+152}$

It should be recognized at the outset that when relaxation techniques are used to solve linear simultaneous equations, accuracy can be improved at any stage by simply carrying arithmetic operations to more digits. Therefore, at an early stage, little is to be gained by finding "exact" multiples of the x's to counterbalance given or resulting residuals. As will be demonstrated in the solution that follows, greater accuracy and ease of solution can be achieved by operating on multiples (by factors of 10^n) of the unknowns and corresponding residuals.

The solution is summarized in Table 5.21. The solution, to the accuracy thus far obtained, is

$$\begin{aligned} x_1 &= +84{,}734 \times 10^{-4} = +8.4734 \\ x_2 &= +115{,}090 \times 10^{-4} = +11.5090 \\ x_3 &= -85{,}665 \times 10^{-4} = -8.5665 \\ x_4 &= -62{,}313 \times 10^{-4} = -6.2313 \end{aligned} \tag{5.110}$$

with residuals as follows:

$$\begin{aligned} R_1 &= -20 \times 10^{-4} = -0.0020 \\ R_2 &= -14 \times 10^{-4} = -0.0014 \\ R_3 &= +10 \times 10^{-4} = +0.0010 \\ R_4 &= -22 \times 10^{-4} = -0.0022 \end{aligned} \tag{5.111}$$

Prob. 5.1: Obtain the solution to the following set of equations by the Gauss, Cholesky, square-root, iteration, and relaxation methods. (Carry the solution to three decimal places.)

Table 5.21 *Solution of Eqs. (5.109) Using Relaxation Table 5.20*

Line†	x_1	x_2	x_3	x_4	R_1	R_2	R_3	R_4	Note
1	$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$x_4 = 0$	−372	−420	−138	+304	
2	+12		−18	+6	−396	+36	−150	+316	
3	+6	+12	+12		+36	+36	−174	+388	
4	−9	+3	−9	−9	−12	+111	−168	−68	
5		−3	+1	−3	−28	+98	−26	−72	
6	−2		+3	−1	−24	+22	−24	−74	
6′	+7	+12	−11	−7	−24	+22	−24	−74	
7‡	+70	+120	−110	−70	−240	+220	−240	−740	$\times 10^{-1}$
8	+15	−5	+15	+15	−160	+95	−250	+20	↓
9		−6	+2	−6	−192	+69	+34	+12	↓
10	+3	+6	+6		+24	+69	+22	+48	↓
11	−2		+3	−1	+28	−7	+24	+46	↓
11′	+86	+115	−84	−62	+28	−7	+24	+46	$\times 10^{-1}$
12	+860	+1,150	−840	−620	+280	−70	+240	+460	$\times 10^{-2}$
13	−9	+3	−9	−9	+232	+5	+246	+4	↓
14	−3	−6	−6		+16	+5	+258	−32	↓
15		+6	−2	+6	+48	+31	−26	−24	↓
15′	+848	+1,153	−857	−623	+48	+31	−26	−24	$\times 10^{-2}$
16	+8,480	+11,530	−8,570	−6,230	+480	+310	−260	−240	$\times 10^{-3}$
17	−7	−14	−14		−24	+310	−232	−324	↓
18	+6	−2	+6	+6	+8	+260	−236	−20	↓
19	−8		+12	−4	+24	−44	−224	−28	↓
20		−6	+2	−6	−8	−70	+60	−36	↓
21	+2		−3	+1	−12	+6	+58	−34	↓
21′	+8,473	+11,508	−8,567	−6,233	−12	+6	+58	−34	$\times 10^{-3}$
22	+84,730	+115,080	−85,670	−62,330	−120	+60	+580	−340	$\times 10^{-4}$
23		+12	−4	+12	−56	+112	+12	−324	↓
24	+6	−2	+6	+6	−24	+62	+8	−20	↓
25	−2		+3	−1	−20	−14	+10	−22	↓
25′	+84,734	+115,090	−85,665	−62,313	−20	−14	+10	−22	$\times 10^{-4}$

† Lines designated by primed numbers summarize solutions obtained to that point.

‡ Since the values of the residuals are small in comparison with the operators listed in Table 5.20, it is desirable to multiply all quantities by a factor of 10 before relaxing further. That is, $f_i(10x_1, 10x_2, 10x_3, 10x_4) = g_i(10R_1, 10R_2, 10R_3, 10R_4)$. The corresponding set is given by line 7.

$$\begin{aligned}
+1x_1 \quad -1x_2 \qquad\qquad\qquad\qquad\qquad &= +1\\
-1x_1 \quad +2x_2 \quad -1x_3 \qquad\qquad\qquad &= +2\\
-1x_2 \quad +3x_3 \quad -1x_4 \qquad &= +3\\
-1x_3 \quad +2x_4 \quad -1x_5 &= +2\\
-1x_4 \quad +1x_5 &= +1
\end{aligned}$$

Prob. 5.2: Solve the problem in Example 4.4 for I_1, I_2, I_3, and I_4 in terms of E, where

$$\begin{aligned}
R_1 &= R & E_1 &= E\\
R_2 &= R & E_2 &= 2E\\
R_3 &= R\\
R_4 &= 2R\\
R_5 &= R\\
R_6 &= 3R
\end{aligned}$$

Prob. 5.3: Using Cholesky's method, solve the following set of equations:

$$\begin{aligned}
6x_1 \quad + 3x_2 \qquad\qquad\qquad\qquad &= +36\\
3x_1 \quad +16x_2 \quad + 5x_3 \qquad\qquad &= +120\\
5x_2 \quad +14x_3 \quad +2x_4 &= -124\\
2x_3 \quad +4x_4 &= +2
\end{aligned}$$

Prob. 5.4: By solving the following 3 × 3 set of equations by the square-root method, verify that no special difficulties are encountered when u_{ii} is the square root of a negative number. (*Hint:* $\sqrt{-a^2} = ai$.)

$$\begin{aligned}
4x_1 \quad + 2x_2 \qquad\qquad &= +2\\
2x_1 \quad -10x_2 \quad +4x_3 &= +6\\
4x_2 \quad +5x_3 &= +5
\end{aligned}$$

Prob. 5.5: If a digital computer does not recognize the difference between $\sqrt{-a^2}$ and $\sqrt{+a^2}$, will a solution by the square-root method be usable? Why?

Prob. 5.6: Solve, by relaxation methods, the following set of equations [*Note:* Eq. (5.21) and Table 5.2]:

$$\begin{aligned}
3x_1 \quad -3x_2 \quad +5x_3 &= +4\\
1x_1 \quad +2x_2 \quad -6x_3 &= +3\\
2x_1 \quad -1x_2 \quad +3x_3 &= +1
\end{aligned}$$

Prob. 5.7: Solve, by any of the presented methods, the following sets of equations:

(*a*)
$$\begin{bmatrix} 0 & +2 & -1 \\ -2 & 0 & -2 \\ +1 & +2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +4 \\ +2 \\ -2 \end{bmatrix}$$

$$D = \begin{vmatrix} 0 & 2 & -1 \\ 0 & 4 & -2 \\ 1 & 2 & 0 \end{vmatrix} = (-1)^{3+1} \begin{vmatrix} 2 & -1 \\ 4 & -2 \end{vmatrix}$$

$$\leftarrow \boxed{D = 0}$$

(*b*)
$$\begin{bmatrix} +4 & 0 & +2 \\ 0 & +4 & 0 \\ +2 & 0 & +4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +8 \\ +2 \\ +4 \end{bmatrix}$$

(*c*)
$$\begin{bmatrix} +10 & +8 & +9 \\ +9 & -7 & +10 \\ -13 & -14 & +12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +35 \\ -25 \\ +14 \end{bmatrix}$$

(*d*)
$$\begin{bmatrix} -3.2 & +4.6 & +4.3 \\ +4.2 & +3.2 & +2.8 \\ +0.6 & -0.8 & +0.4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} +2.0 \\ +4.4 \\ -0.3 \end{bmatrix}$$

(*e*)
$$\begin{bmatrix} +2 & +3 & 0 & +4 \\ +3 & +20 & +6 & -4 \\ 0 & +3 & +2 & 0 \\ +4 & -5 & 0 & +3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ +2 \\ +1 \\ 0 \end{bmatrix}$$

(*f*)
$$\begin{bmatrix} +10 & 0 & +2 & -1 \\ +2 & +12 & +3 & -4 \\ +1 & -4 & +18 & +6 \\ 0 & +3 & -3 & -12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ +1 \\ -1 \end{bmatrix}$$

(*g*)
$$\begin{bmatrix} +1.8 & +1.2 & 0 & 0 \\ +1.2 & +2.4 & +1.2 & 0 \\ 0 & +1.2 & +2.4 & +1.2 \\ 0 & 0 & +1.2 & +3.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} +2.8 \\ +1.0 \\ +1.0 \\ +1.6 \end{bmatrix}$$

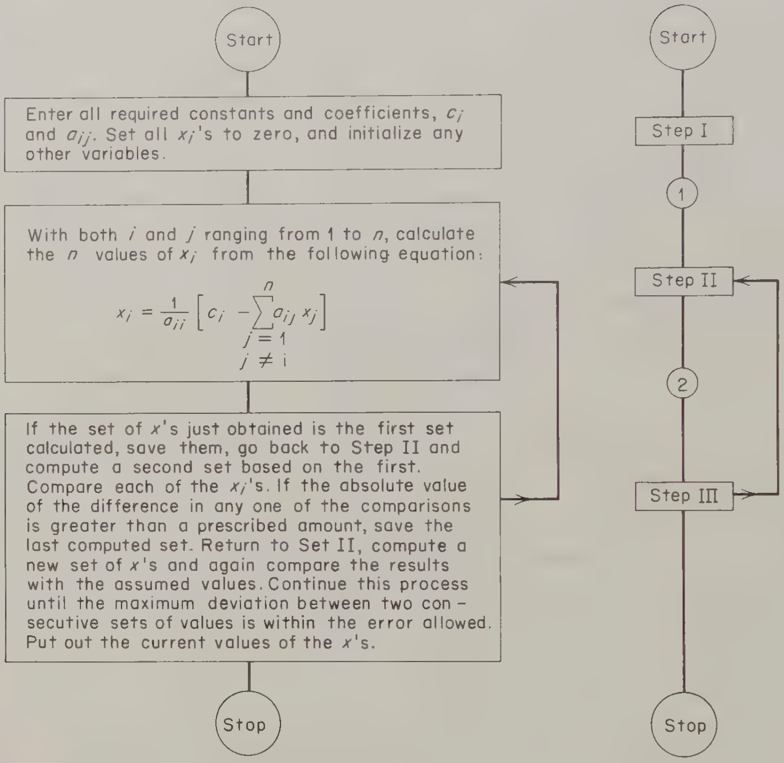
Prob. 5.8: Solve Example 4.1 for values of $k = 0$, $\pi/2$, and π.

Prob. 5.9: Show that the product of the elements ℓ_{ii} is equivalent to $|A|$.

5.6 *Computer Solutions*

In selecting a method of solution for digital-computer programming, two factors are normally considered. The first of these is concerned with the ease of programming; the more repetitive the calculational process, the easier is the programming. Secondly, the running time of solution is a major factor. For smaller computers, a third consideration, and the one that may be the controlling factor, is storage capacity.

Flow diagrams for two different types of solutions will be illustrated in



Enter all required constants and coefficients, c_i and a_{ij}. Set all x_i's to zero, and initialize any other variables.

With both i and j ranging from 1 to n, calculate the n values of x_i from the following equation:

$$x_i = \frac{1}{a_{ii}}\left[c_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j\right]$$

If the set of x's just obtained is the first set calculated, save them, go back to Step II and compute a second set based on the first. Compare each of the x_i's. If the absolute value of the difference in any one of the comparisons is greater than a prescribed amount, save the last computed set. Return to Set II, compute a new set of x's and again compare the results with the assumed values. Continue this process until the maximum deviation between two consecutive sets of values is within the error allowed. Put out the current values of the x's.

Gauss-Seidel method

Figure 5.5

this section. The ones that have been chosen for discussion are the Gauss-Seidel method (iteration) and the Cholesky method (elimination).

5.6.1 Gauss-Seidel Iteration

GOOD FAST PROGRAM

As was demonstrated in Sec. 5.4.2, the solution to sets of linear equations by iteration requires the repeated execution of a given elementary computational process. Moreover, the results of any one set of calculations depend on the last preceding set of calculated solutions. Such systems are ideally suited to flow diagramming, as will be apparent from the following discussion.

In broad outline, Fig. 5.5 is a flow diagram that could be used to obtain the solution to sets of simultaneous linear algebraic equations by using the Gauss-Seidel (iteration) method. A more detailed consideration of the specific steps necessary to carry out these instructions is given in Figs. 5.6 to 5.8. The complete flow diagram is shown in Fig. 5.9.

STEP I

The flow diagram in Fig. 5.6 will enter the data and initialize the variables. With regard to the notes indicated in Fig. 5.6:

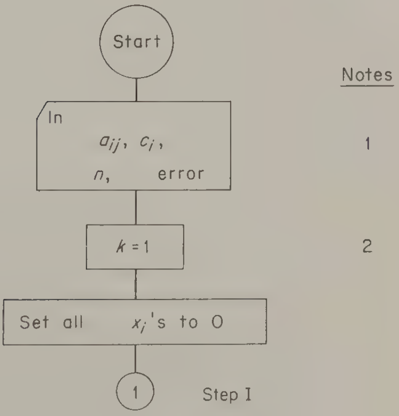
1. The coefficients of x, that is, a_{ij}, the values of c_i, and the number n are read into the computer. "Error" is a term which represents the maximum allowable difference between corresponding values in any two consecutive sets of computed x_i's. "Error," then, represents the precision to which the values of x are to be computed. For example, if the values of x are desired to five decimal places, the prescribed error is 0.000005.
2. k is a constant used to identify the first calculated values of the x's. This is needed since the first set of values is to be saved without making comparisons, whereas each subsequent set must first be compared with the set immediately preceding before being used. k will therefore need to be changed after the first set of values has been obtained.

STEP II

With both i and j ranging from 1 to n, calculate the n values of x_i from the following equation:

$$x_i = \frac{1}{a_{ii}} \left(c_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right) \tag{5.112}$$

Figure 5.6



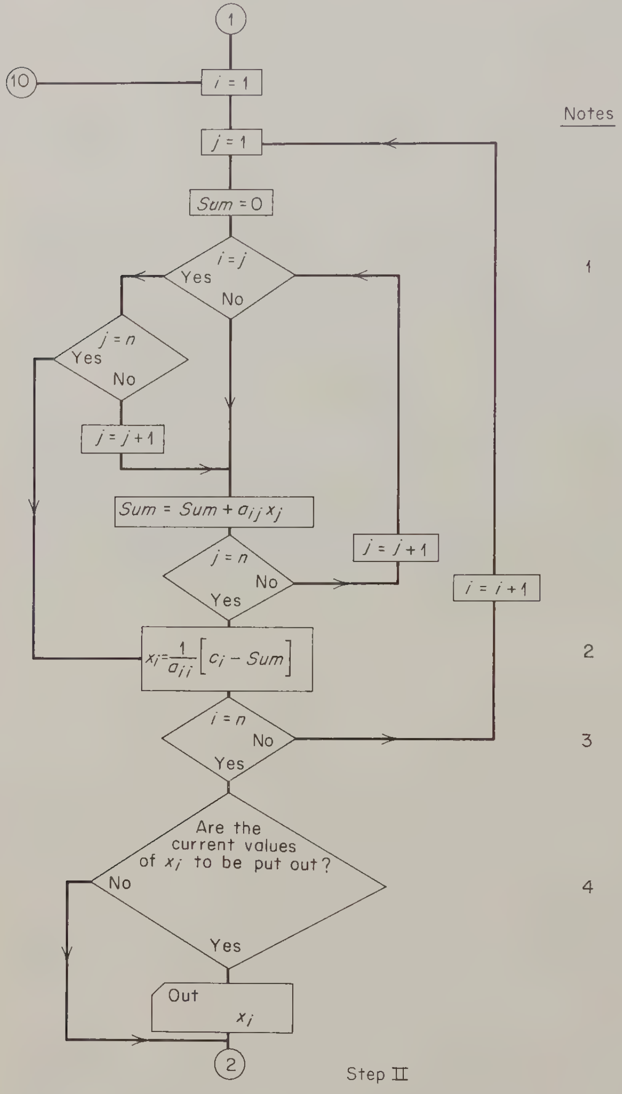


Figure 5.7

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

This part of the flow diagram is shown in Fig. 5.7, with notes as follows:

1. All values of a_{ij} and x_j are to be used in the calculation of any particular x_i except x_i itself. The computational block noted as ① is used to identify that particular entry to be excluded. If i is equal to j, that value of j is skipped and the next value used. However,

before this is done, a check is required to determine whether j has become as large as n. If it has, no additional sums are accumulated. If not, j is incremented by 1.

2. A particular value of x_i is calculated at this point. *Sum* contains the products of all current values of a_{ij} and x_j (with the exception of the ith terms).
3. At this stage, it is necessary to determine whether all n of the values of x have been calculated. If they have not, i is incremented by 1, j is reset to 1, and the process is repeated for a new x_i. If i does equal n, all the x_i's have been determined and computation proceeds to the next step.
4. If the rate of convergence to a final set of x's is slow, it may be desirable to have the computer put out at any time the most recent set of values. Manual control of most computers, through various sensing mechanisms, can provide this feature.

STEP III

If the set of x's just obtained is the first set calculated, save them, go back to Step II, compute a new set of x's, and again compare the results with the assumed values. Continue this process until the maximum deviation between two consecutive sets of values is within the error allowed. Put out the current values of the x's.

This part of the flow diagram is shown in Fig. 5.8, with notes as follows:

1. If at any time in the computation the rate of convergence is found to be too slow, the process can be accelerated by jumping ahead to new values of x_i. This block allows for the introduction of such values. The mechanics of such introduction is much the same as for an intermediate output.
2. At this stage in the calculations, it is necessary to determine whether the computed set of values of x_i is the first (that is, $k = 1$) or some later set. (It will be recalled that in Step I the variable k was introduced to facilitate comparison between the first and second sets of computed values.) Immediately after identifying the first set, the value of k is changed to zero, and subsequent exits are through the path marked "No." Each of the values of x_i is placed in a temporary location called sx_i. When a complete set has been stored, the entire process is repeated (⑩) and a new set is computed.
3. Each set of values of x is compared with its corresponding values in the previous set, that is, sx_i, to determine whether any of the individual differences exceed in magnitude the initially prescribed error.

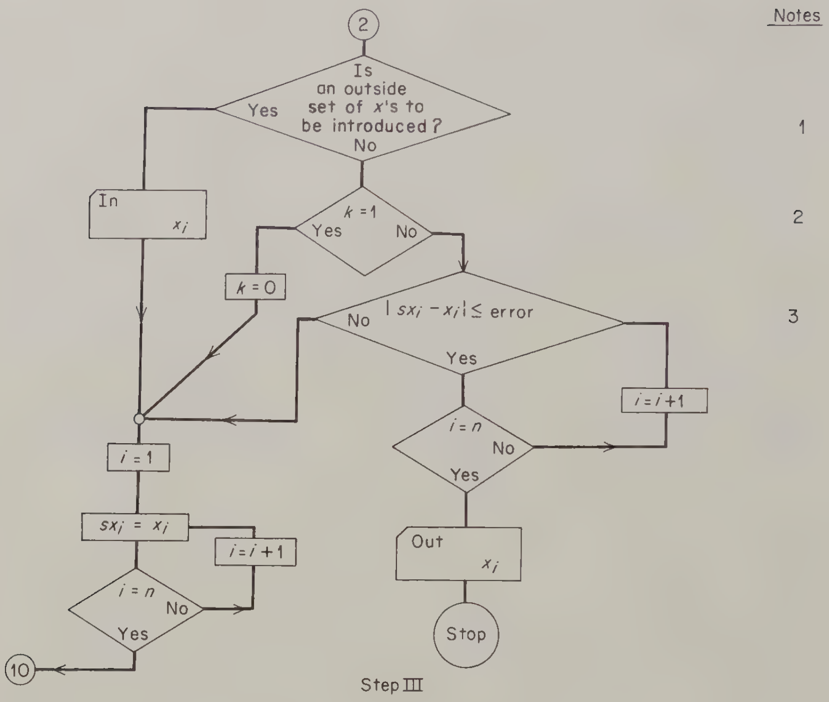


Figure 5.8

> If for any of the x_i's the difference is found to be greater than that allowed, the last computed set is saved in the sx_i location and a new set of x_i's is computed. When all the values of x_i satisfy the error test, these values are put out as the final answer.

Figure 5.9 is a complete flow diagram of the Gauss-Seidel method for solving simultaneous linear algebraic equations.

5.6.2 Cholesky's Method

As is true of all the elimination procedures, Cholesky's method reduces the given set of equations $[A][X] = [C]$ to the upper unit triangular form $[T][X] = [K]$. Back substitution then can be used to produce the required values of x.

Reduction to the upper unit triangular matrix is accomplished by use of a lower triangular matrix $[L]$. Since one of the computations that must be carried out is a division by ℓ_{ii} [see Eq. (5.62)], the major diagonal element of $[L]$, it is necessary that this element be other than zero. If a

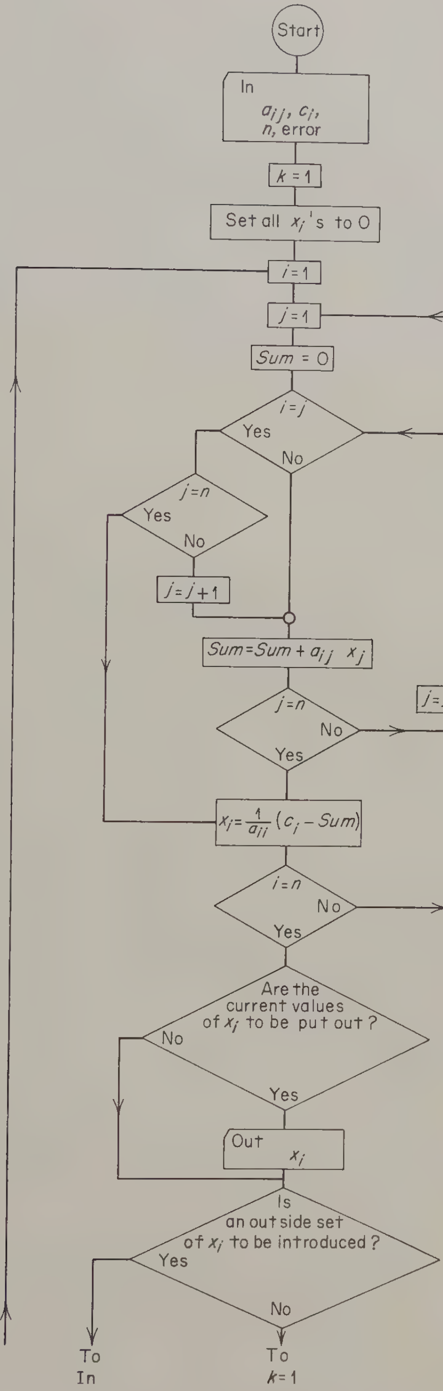


Figure 5.9 Flow diagram for Gauss-Seidel method—solution of simultaneous linear algebraic equations.

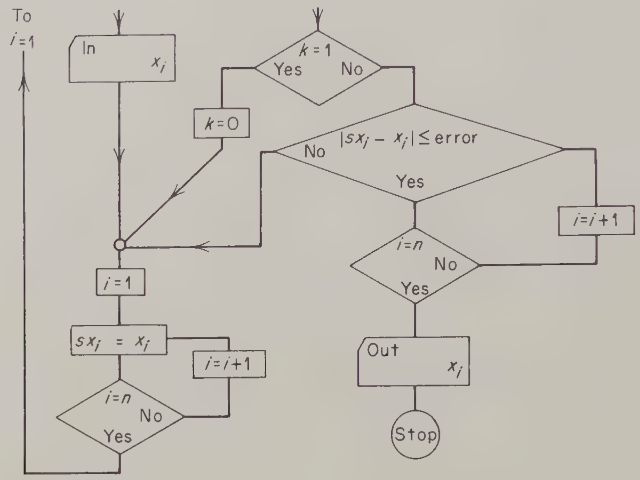


Figure 5.9 (*Continued*)

zero is found to exist, rows must be rearranged in such a way as to eliminate the zero before solution proceeds.

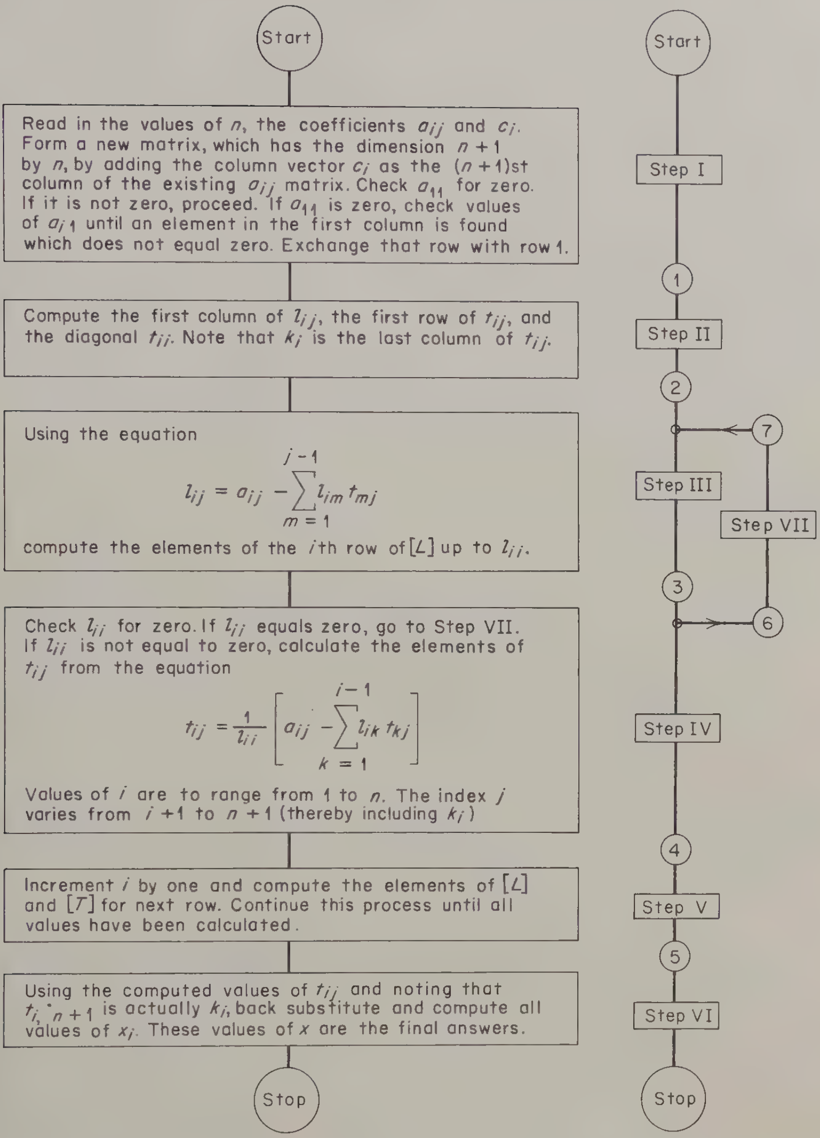
A very general flow diagram for the solution to this problem is shown in Fig. 5.10. Detailed flow diagrams of the various segments of the general program are shown in Figs. 5.11 through 5.17, and a complete diagram is shown in Fig. 5.18.

STEP I

Read in the values of n, the coefficients a_{ij}, and c_i. Form a new matrix, which has the dimensions $n + 1$ by n, by adding the column vector c_i as the $(n + 1)$st column of the existing a_{ij} matrix. Check a_{11} for zero. If it is not, proceed. If a_{11} is zero, check values of a_{i1} until an element in the first column is found which does not equal zero. Exchange that row with row 1.

A detailed flow diagram is shown in Fig. 5.11 (see p. 107). The notes indicated in the figure have the following significance:

1. The value of n read in at this point is the size of the set of equations to be solved. a_{ij} and c_i are the coefficients of the unknown x's and the constants, respectively.
2. The coefficient matrix is augmented by adding to it the column vector of constants c_i. These are placed in the $(n + 1)$st column position.
3. In calculating the elements of the first row of $[T]$, it is necessary to divide by the value of a_{11}. Thus, a_{11} cannot equal zero. If it


Start
Read in the values of n, the coefficients a_{ij} and c_i. Form a new matrix, which has the dimension $n+1$ by n, by adding the column vector c_i as the $(n+1)$st column of the existing a_{ij} matrix. Check a_{11} for zero. If it is not zero, proceed. If a_{11} is zero, check values of a_{i1} until an element in the first column is found which does not equal zero. Exchange that row with row 1.
Compute the first column of l_{ij}, the first row of t_{ij}, and the diagonal t_{ii}. Note that k_i is the last column of t_{ij}.
Using the equation $l_{ij} = a_{ij} - \sum_{m=1}^{j-1} l_{im} t_{mj}$ compute the elements of the ith row of $[L]$ up to l_{ii}.
Check l_{ii} for zero. If l_{ii} equals zero, go to Step VII. If l_{ii} is not equal to zero, calculate the elements of t_{ij} from the equation $t_{ij} = \frac{1}{l_{ii}} \left[a_{ij} - \sum_{k=1}^{i-1} l_{ik} t_{kj} \right]$ Values of i are to range from 1 to n. The index j varies from $i+1$ to $n+1$ (thereby including k_i)
Increment i by one and compute the elements of $[L]$ and $[T]$ for next row. Continue this process until all values have been calculated.
Using the computed values of t_{ij} and noting that $t_{i,n+1}$ is actually k_i, back substitute and compute all values of x_i. These values of x are the final answers.
Stop
Start
Step I
1
Step II
2
7
Step III
Step VII
3
6
Step IV
4
Step V
5
Step VI
Stop

(a)

Figure 5.10

has a value other than zero, the program proceeds. If, however, a_{11} is found to be equal to zero, a_{21} is checked to see whether it equals zero. Then a_{31} is checked, and so on until a value of a_{i1} is found which is suitable. If a nonzero a_{i1} cannot be found, the problem stops.

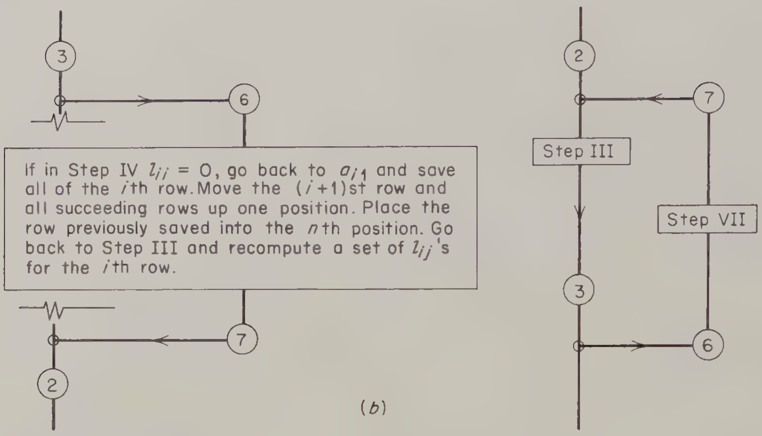

3
6
If in Step IV l_{ii} = 0, go back to a_{i1} and save all of the ith row. Move the $(i+1)$st row and all succeeding rows up one position. Place the row previously saved into the nth position. Go back to Step III and recompute a set of l_{ij}'s for the ith row.
7
2
(b)
2
7
Step III
Step VII
3
6

Figure 5.10 *(Continued)*

4. This check is to determine the current value of i and thereby to identify the row for which a_{i1} was not zero. If the current value is 1, the problem proceeds to Step II. If the current value of i is not 1, the ith row is interchanged with the first row.

STEP II

Compute the first column of ℓ_{ij}, the first row of t_{ij}, and the diagonal t_{ii}. Note that k_i is the last column of t_{ij}.

The following are noted in Fig. 5.12:

1. In Cholesky's method, $[L]$ is a lower triangular matrix, the first column of which is equal to the first column of the matrix $[A]$.
2. $[T]$ is an upper unit triangular matrix. This step in the flow diagram sets the diagonal elements of t_{ij} (that is, t_{ii}) equal to unity.
3. The first row of t_{ij} equals the first row of a_{ij} divided by a_{11}. However, this step is begun at $j = 2$, since t_{11} has already been set to unity. When all values of t_{1j} have been calculated, including $t_{1,n+1}$, computation proceeds to Step III.

STEP III

Using the equation

$$\ell_{ij} = a_{ij} - \sum_{m=1}^{j-1} \ell_{im} t_{mj}$$

compute the elements of the ith row of $[L]$ up to ℓ_{ii}.

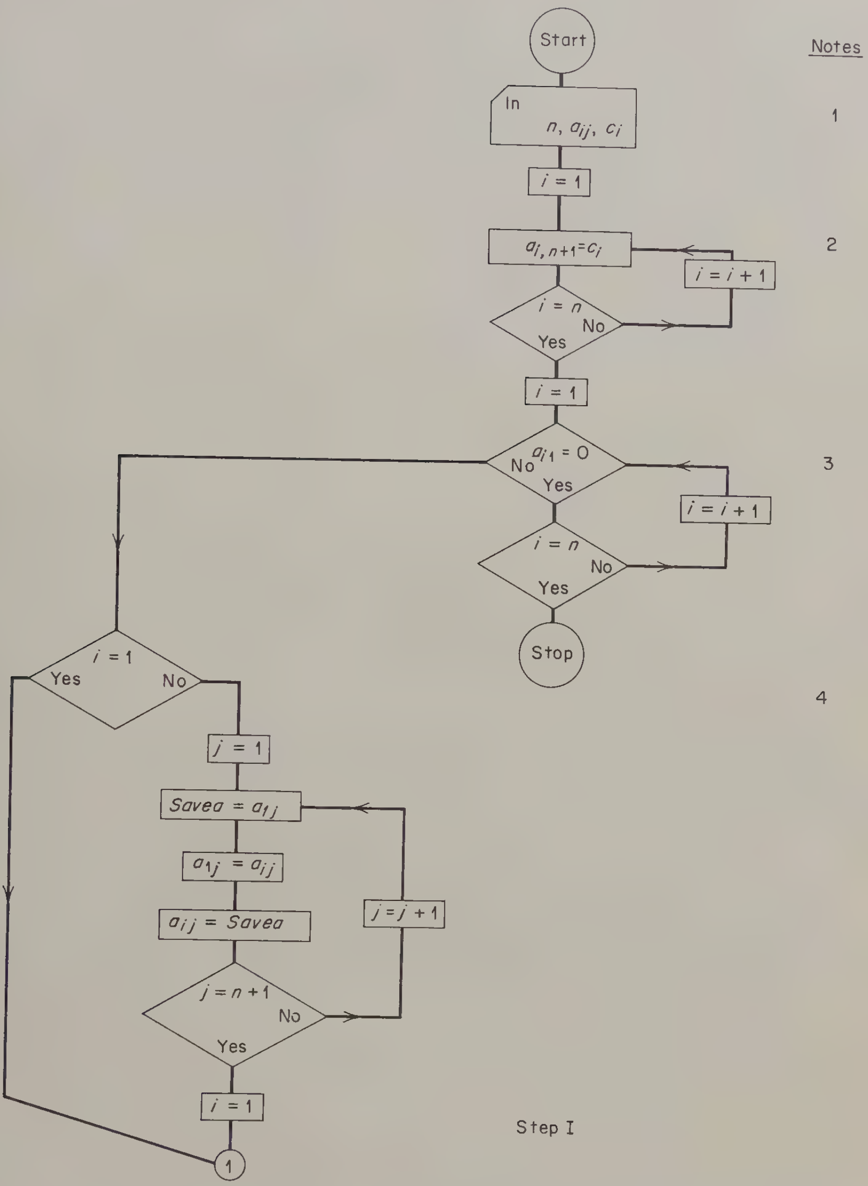

Notes
Start
In
n, a_{ij}, c_i
1
$i = 1$
$a_{i,n+1} = c_i$
2
$i = i + 1$
$i = n$
No
Yes
$i = 1$
$a_{i1} = 0$
No
Yes
3
$i = i + 1$
$i = n$
No
Yes
Stop
$i = 1$
Yes
No
4
$j = 1$
$Savea = a_{1j}$
$a_{1j} = a_{ij}$
$a_{ij} = Savea$
$j = j + 1$
$j = n + 1$
No
Yes
$i = 1$
1
Step I

Figure 5.11

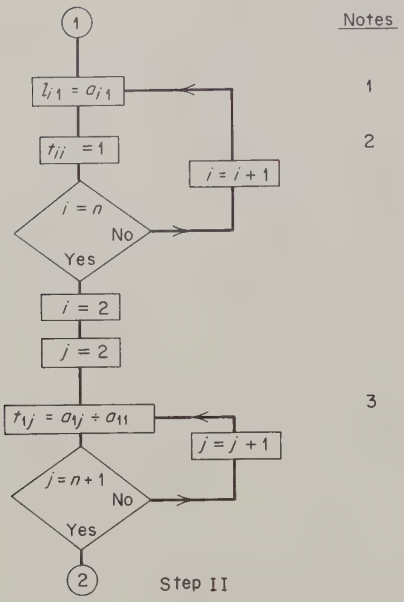


Figure 5.12

That part of the total flow diagram which causes this step to be accomplished is shown in Fig. 5.13. Each of the steps is self-evident, with the possible exception of the ones noted.

1. Index *ik* is a variable used to keep account of the current value of *i*.
2. Since j is initially set at 2, ℓ_{i2} is the first element of ℓ_{ij} to be determined. (Note that ℓ_{i1} has already been defined.) When j has been increased to i, the problem passes to Step IV.

In each row, values of ℓ_{ij} beyond ℓ_{ii} are zero.

STEP IV

Check ℓ_{ii} for zero. If ℓ_{ii} equals zero, go to Step VII. If ℓ_{ii} is not equal to zero, calculate the elements of t_{ij} from the equation

$$t_{ij} = \frac{1}{\ell_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} \ell_{ik}t_{kj}\right)$$

Values of i are to range from 1 to n. j varies from $i + 1$ to $n + 1$ (thereby including k_i).

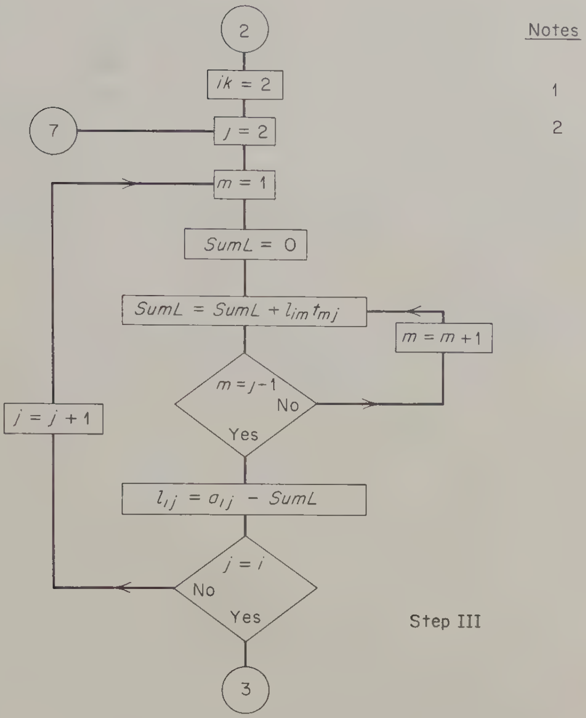


Figure 5.13

With regard to the notes indicated in Fig. 5.14:

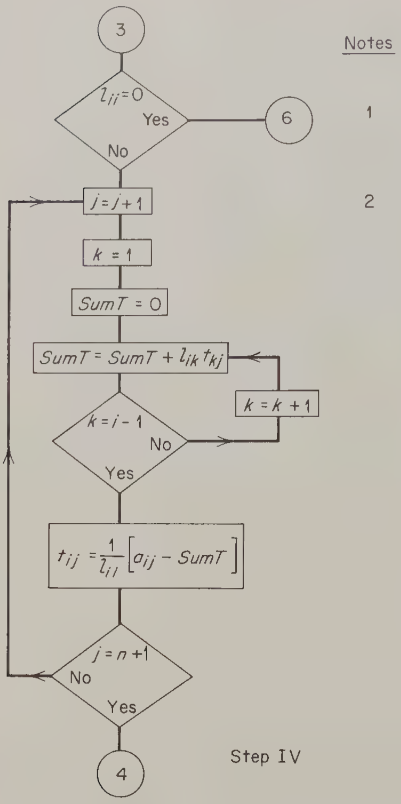
1. Since at this stage division by ℓ_{ii} is required, it is necessary to check whether ℓ_{ii} equals zero. If not, the computation proceeds. If, however, it is equal to zero, control passes through exit ⑥ to a special routine, the purpose of which is to eliminate that zero (Step VII).
2. When entering this part of the flow diagram, the value of j was the same as i. Since t_{ii} has already been defined, j is set at $j + 1$, and $t_{i,i+1}$ is the first element computed. In succession, each element for that row is found, up to $t_{i,n+1}$ (which is actually k_i).

STEP V

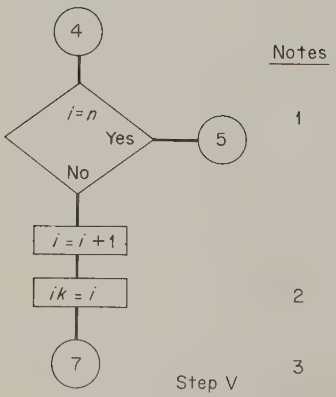
Increment i by 1 and compute the elements of $[L]$ and $[T]$ for the next row. Continue this process until all values have been calculated.

Regarding the notes in Fig. 5.15:

1. At this stage, only one row of ℓ_{ij} and t_{ij} have been found. This part of the flow diagram causes each of the remaining rows to $[L]$ and $[T]$ to be calculated.



Step IV



Step V

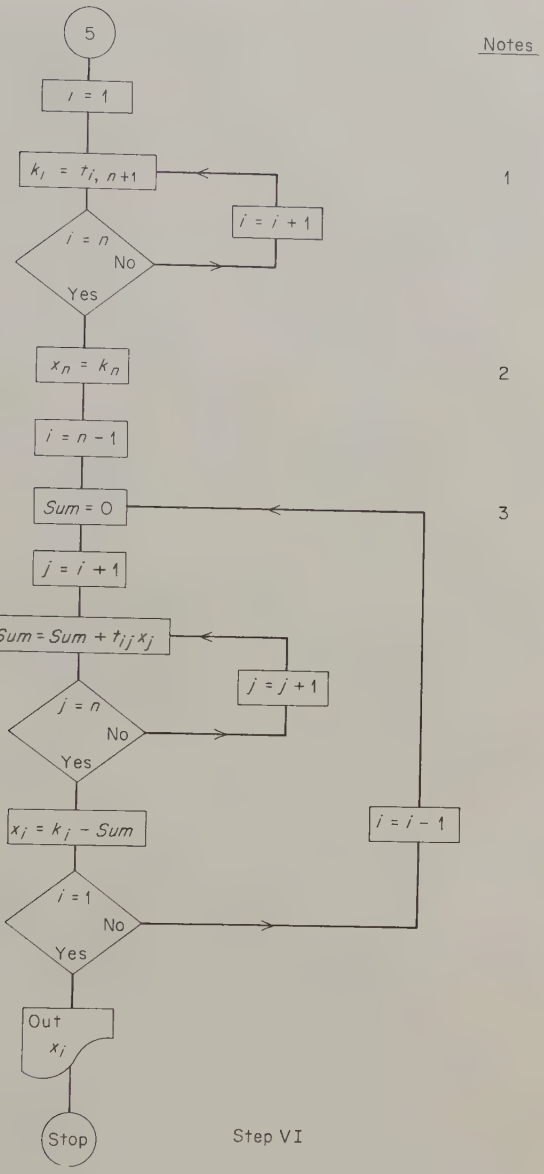


Figure 5.16

2. *ik* is a variable used to keep account of the current value of i. It is not used unless ℓ_{ii} is found to equal zero, in which case the rows must be interchanged.
3. ⑦ is the point in Step III (see Fig. 5.13) to which the problem returns to compute a new row of $[L]$ and $[T]$.

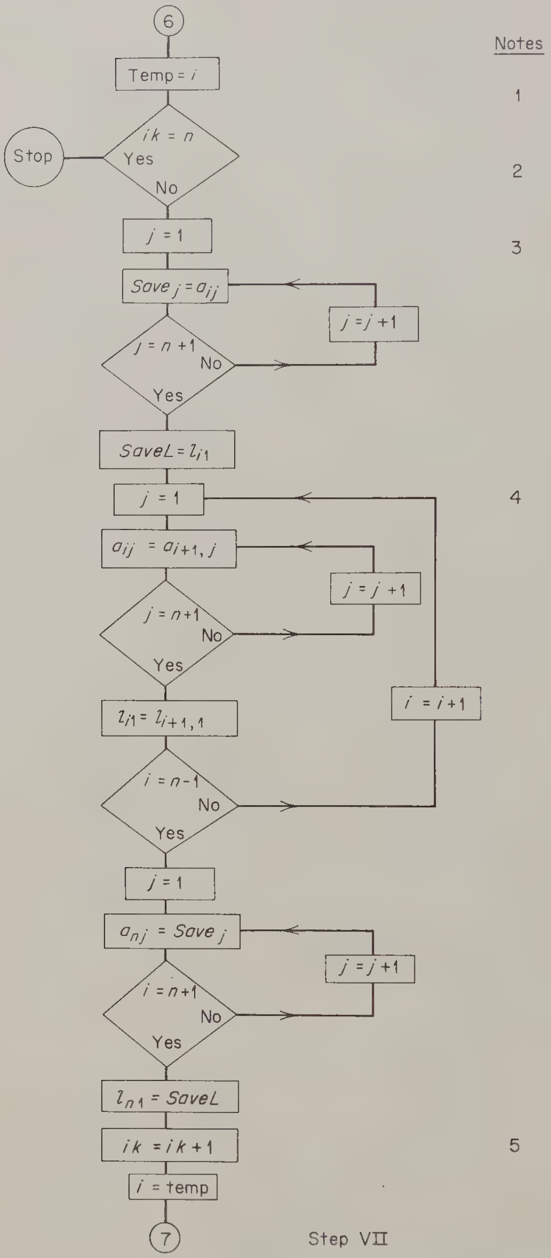

6
Notes
Temp = i
1
ik = n
Yes
No
Stop
2
j = 1
3
Save j = a_ij
j = j + 1
j = n + 1
No
Yes
SaveL = l_i1
j = 1
4
a_ij = a_i+1, j
j = j + 1
j = n+1
No
Yes
i = i + 1
l_i1 = l_i+1, 1
i = n - 1
No
Yes
j = 1
a_nj = Save j
j = j + 1
i = n+1
No
Yes
l_n1 = SaveL
ik = ik + 1
5
i = temp
7
Step VII

Figure 5.17

STEP VI

Using the computed values of t_{ij} and noting that $t_{i,n+1}$ is actually k_i, back substitute and compute all values of x_i. These values of x are the final answers.

The notes in Fig. 5.16 correspond to the following:

1. At this stage, each element in the last column of the augmented matrix t_{ij} is placed into a special location called k_i. These values are used in the back-substitution routine with the equations $[T][X] = [K]$.
2. Since t_{ij} is an upper unit triangular matrix, the nth value of k_i is equal to the value of x_n.
3. Each value of x_i, beginning with x_{n-1} and continuing to x_1, is calculated by back substitution. When all values of x have been found, they are given as output and the problem is complete.

STEP VII

If, in Step IV, $\ell_{ii} = 0$, go back to a_{i1} and save a of the ith row. Move the $(i + 1)$st row and all succeeding rows up one position. Place the row previously saved into the nth position. Go back to Step III and recompute a set of ℓ_{ij}'s for the new ith row.

Figure 5.17 is a flow diagram which will result in the required row rearrangement. The notes indicated in the figure correspond to the following:

1. The value of i corresponding to $\ell_{ii} = 0$ must be saved for future use. *Temp* is the location in which i is to be saved.
2. If, at this stage, ik is found to be equal to n and $\ell_{nn} = 0$, no rearrangement of rows can be accomplished and the problem must stop. If ik is not equal to n, rows can possibly be rearranged.
3. Beginning with a_{i1}, i being the row in which ℓ_{ii} was found to be equal to zero, each a_{ij} is saved in a location called $Save_j$.
4. The $(i + 1)$st row is moved up to take the place of the ith row. Each succeeding row is also moved up one row. The nth row is replaced by the (former) ith row, thus completing the rearrangement.
5. The value of ik must be increased by 1 in order that the number of rearrangements is available at all times. Variable i is reset to *Temp*, and a new set of values of ℓ_{ij} is computed. If the new ℓ_{ii} is again found to be zero, rows are again rearranged. Since ik is increased by 1 each time a rearrangement is required, the problem will stop if any row is moved into the ith row which has already been tried.

Figure 5.18 (pp. 115–117) is a complete flow diagram for the evaluation of simultaneous linear algebraic equations using Cholesky's method.

Prob. 5.10: Construct a flow diagram for the evaluation of n simultaneous linear algebraic equations by use of Cramer's rule.

Prob. 5.11: Assuming that a relaxation table is defined, develop a flow diagram for the evaluation of n simultaneous linear algebraic equations by use of relaxation methods.

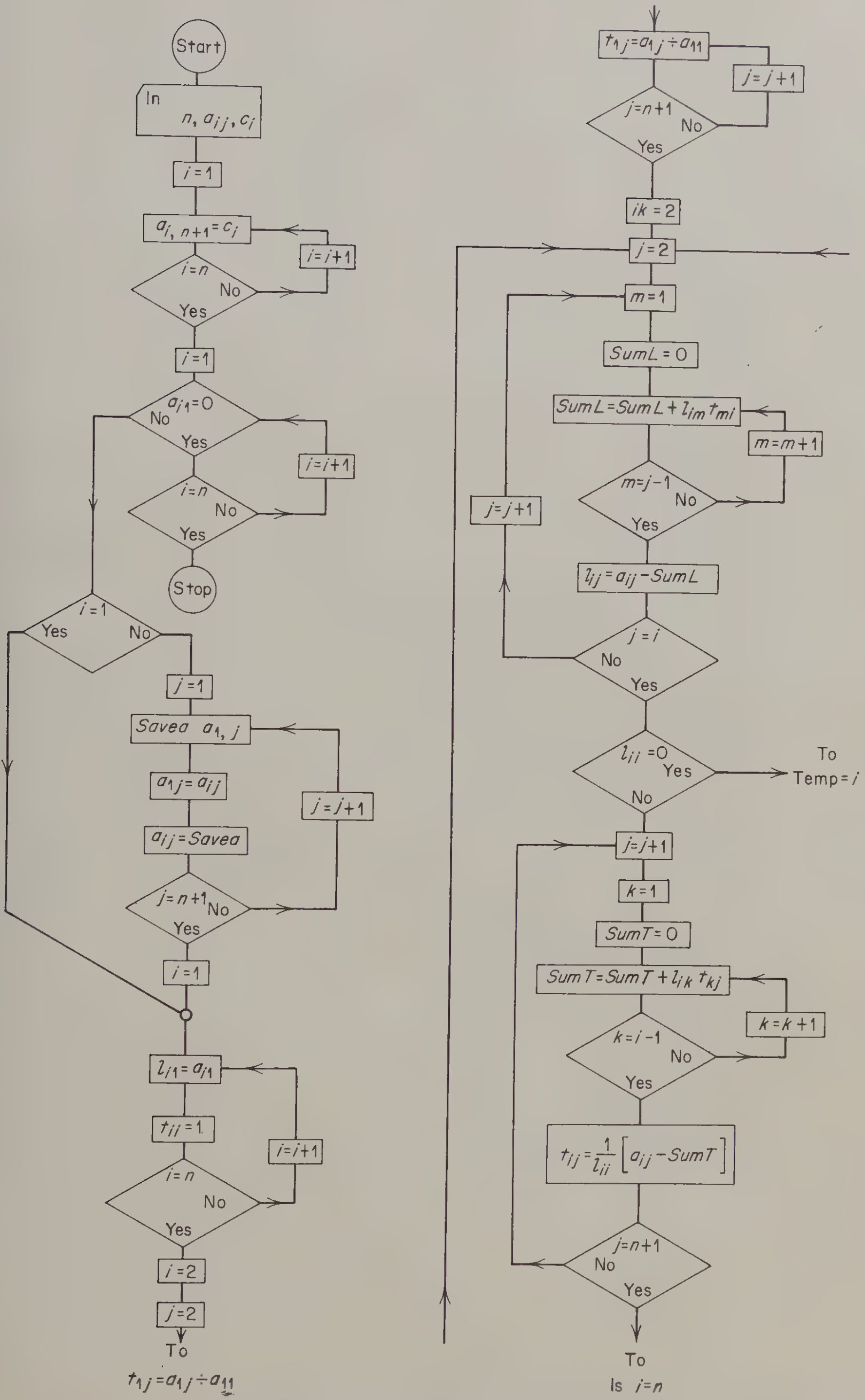


Figure 5.18 Flow diagram for Cholesky's method—solution of simultaneous linear algebraic equations.

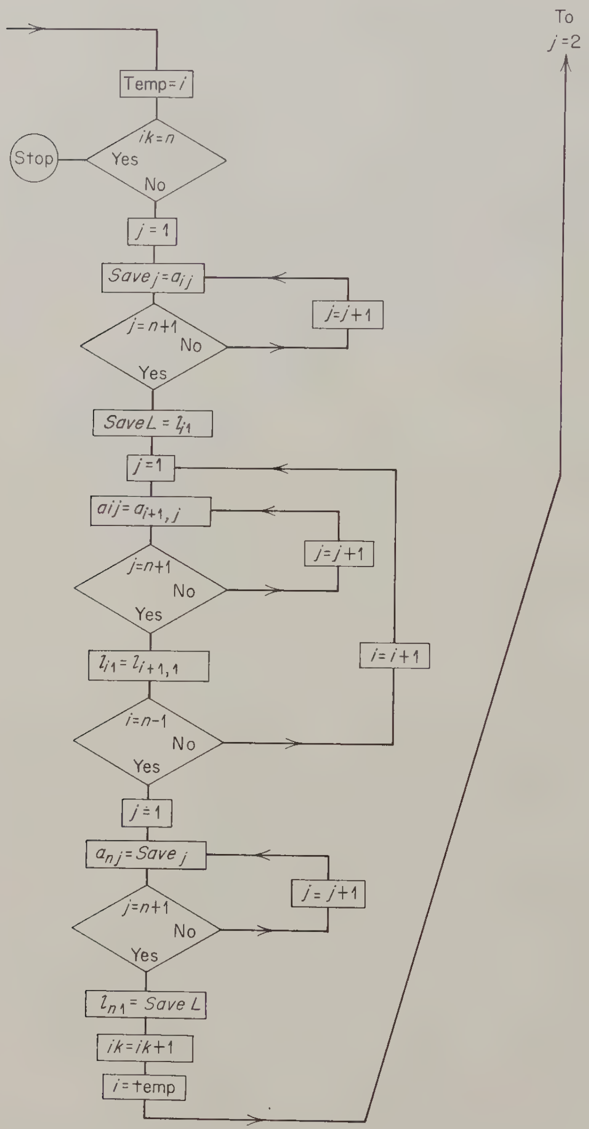


Figure 5.18 *(Continued)*

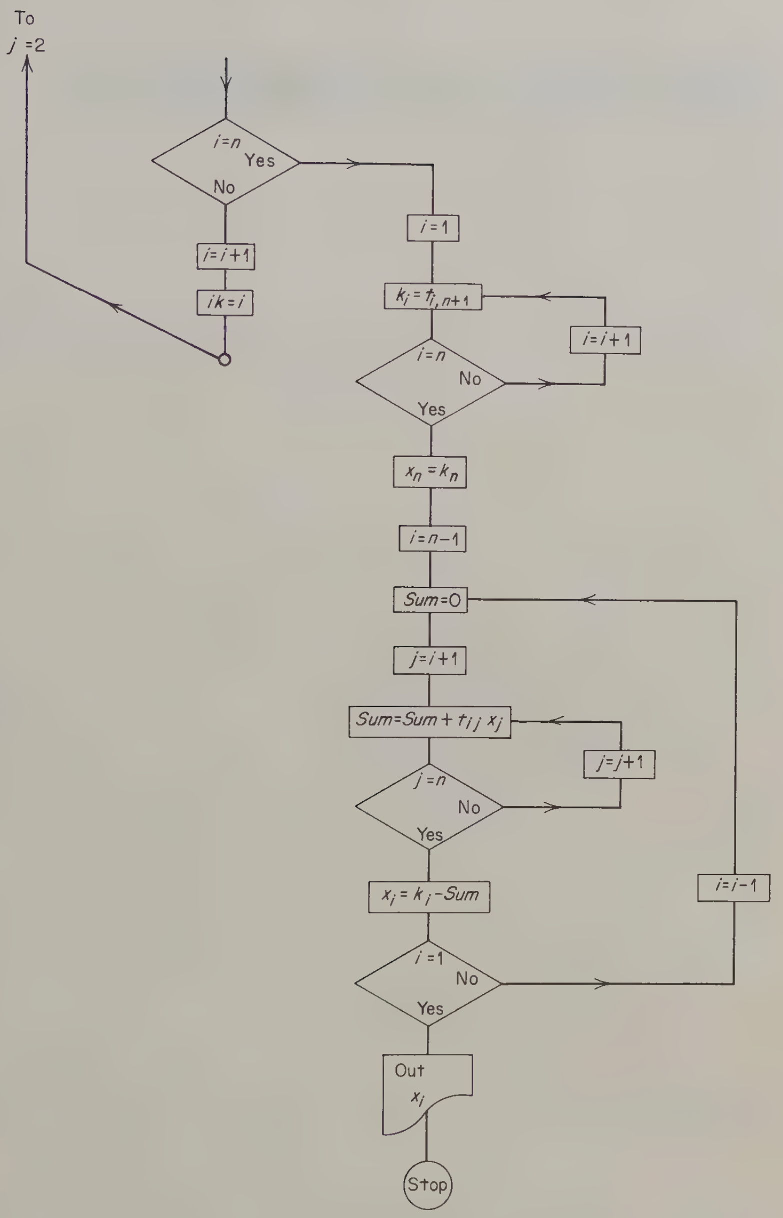


Figure 5.18 *(Continued)*

the inverse problem

6

6.1 Introduction

This chapter is an extension of the preceding discussion of simultaneous linear algebraic equations to include those situations where the given equations are related not to a column vector of constants but rather to another variable. The more general set of equations under consideration is, then,

$$[A][X] = [E][Y] \tag{6.1}$$

where $[A]$ and $[E]$ are coefficient matrices and $[X]$ and $[Y]$ represent column vectors. In expanded form for a 3×3 set, Eqs. (6.1) can be written as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= e_{11}y_1 + e_{12}y_2 + e_{13}y_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= e_{21}y_1 + e_{22}y_2 + e_{23}y_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= e_{31}y_1 + e_{32}y_2 + e_{33}y_3 \end{aligned} \tag{6.2}$$

In most cases, engineering problems and problems in applied science can be formulated so that each of the y's is expressed explicitly in terms of one of the appropriate $a_{ij}x_j$ equations. That is, $[E]$, the general coefficient matrix of the column vector $[Y]$, is the identity matrix $[I]$:

$$[E] = [I] \tag{6.3}$$

For this case, the general set of equations to be considered is given as Eq. (6.4) or (6.5):

$$[A][X] = [I][Y] \tag{6.4}$$

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = y_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = y_n \end{array} \tag{6.5}$$

The *inverse problem* is concerned with the manipulation of Eq. (6.4) or (6.5) so that the x's are explicitly stated in terms of the y's. Given

$$[A][X] = [Y] \tag{6.6}$$

the equations being sought are

$$[X] = [B][Y] = [A^{-1}][Y] \tag{6.7}$$

where $[A^{-1}]$ is called A *inverse*. The problem is to determine the relationship between the original a_{ij}'s and the b_{ij}'s, which define the inverse matrix.

If the value of $[Y]$ specified by Eq. (6.6) is substituted into the defined equations (6.7), it is noted that the matrix product of $[A]$ and $[A^{-1}]$ is the identity matrix $[I]$.

$$[I][X] = [X] = [A^{-1}][Y] = [A^{-1}][A][X] \tag{6.8}$$

or

$$[A^{-1}][A] = [A][A^{-1}] = [I] \tag{6.9}$$

Extensive use of this property will be made in certain of the following methods for matrix inversion.

6.2 *Cramer's Rule*

METHOD #1

One of the most straightforward methods for inverting a matrix is a direct consequence of Cramer's rule (see Sec. 5.2). If it is assumed that the jth value of x is being sought, Eq. (5.17) indicates the following:

$$x_j = \frac{D_j}{D} \qquad (j = 1, 2, \ldots, n) \tag{6.10}$$

In this equation, D is the value of the determinant of the original coefficient matrix a_{ij}, and D_j is the value of the same determinant with the jth column replaced by the column of constants (in this case the y_i's). If expansion by cofactors along this jth column is assumed, Eq. (6.10) can be written in the summation form

$$x_j = \frac{1}{D} \sum_{i=1}^{n} y_i A_{ij} \qquad (j = 1, 2, \ldots, n) \tag{6.11}$$

or

$$x_j = \sum_{i=1}^{n} \frac{A_{ij}}{D} y_i \qquad (j = 1, 2, \ldots, n) \tag{6.12}$$

where the A_{ij}'s are the cofactors of the jth column elements of the original matrix. Clearly, Eq. (6.12) is in the general form desired; however, it should be noted that the i and j in Eq. (6.12) correspond to the row and column position in the original matrix and not to the position desired in the inverse set. This is perhaps better seen by an expansion of Eq. (6.12).

$$\begin{aligned}
x_1 &= \frac{A_{11}}{D} y_1 + \frac{A_{21}}{D} y_2 + \frac{A_{31}}{D} y_3 + \cdots + \frac{A_{n1}}{D} y_n \\
x_2 &= \frac{A_{12}}{D} y_1 + \frac{A_{22}}{D} y_2 + \frac{A_{32}}{D} y_3 + \cdots + \frac{A_{n2}}{D} y_n \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
x_n &= \frac{A_{1n}}{D} y_1 + \frac{A_{2n}}{D} y_2 + \frac{A_{3n}}{D} y_3 + \cdots + \frac{A_{nn}}{D} y_n
\end{aligned} \tag{6.13a}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \frac{1}{D} \begin{bmatrix} A_{11} & A_{21} & A_{31} & \cdots & A_{n1} \\ A_{12} & A_{22} & A_{32} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \\ & & & & \\ A_{1n} & A_{2n} & A_{3n} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \tag{6.13b}$$

In other words, the inverse matrix can be obtained by writing the transpose of the matrix composed of the reduced cofactors (cofactors divided by D) of the original set. In expanded form, the inverse of $[A]$ is given by

$$[A^{-1}] = \frac{1}{D} \begin{bmatrix} A_{11} & A_{21} & A_{31} & \cdots & A_{n1} \\ A_{12} & A_{22} & A_{32} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & A_{3n} & \cdots & A_{nn} \end{bmatrix} \tag{6.14}$$

The A_{ij}'s are the cofactors of the original a_{ij}'s.

To illustrate matrix inversion using this procedure, consider the following two simultaneous equations:

$$\begin{aligned} 2x_1 + x_2 &= y_1 \\ 2x_1 + 3x_2 &= y_2 \end{aligned} \tag{6.15}$$

$A_{ij} = (-1)^{i+j} M_{ij}$

or

$$\begin{bmatrix} +2 & +1 \\ +2 & +3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{6.16}$$

Summarizing the various quantities required for a solution,

$$\begin{gathered} A = \begin{bmatrix} +2 & +1 \\ +2 & +3 \end{bmatrix} \qquad D = \begin{vmatrix} +2 & +1 \\ +2 & +3 \end{vmatrix} = +6 - 2 = +4 \\ A_{11} = +3 \qquad A_{12} = -2 \qquad A_{21} = -1 \qquad A_{22} = +2 \end{gathered} \tag{6.17}$$

The inverse is, therefore,

$$[A^{-1}] = \begin{bmatrix} +0.75 & -0.25 \\ -0.50 & +0.50 \end{bmatrix} \tag{6.18}$$

As a check on this result, the matrix product of $[A]$ and $[A^{-1}]$ should equal the identity matrix:

$$[A][A^{-1}] \overset{(?)}{=} [I]$$

$$\begin{aligned} &\begin{bmatrix} +2 & +1 \\ +2 & +3 \end{bmatrix} \begin{bmatrix} +0.75 & -0.25 \\ -0.50 & +0.50 \end{bmatrix} \\ &\qquad = \begin{bmatrix} 1.50 - 0.50 & -0.50 + 0.50 \\ 1.50 - 1.50 & -0.50 + 1.50 \end{bmatrix} = \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \end{aligned} \tag{6.19}$$

A further verification that this is the correct solution can be obtained by a direct-elimination solution of Eqs. (6.15). Subtracting the second of these equations from the first,

$$-2x_2 = y_1 - y_2$$

or

$$x_2 = -0.50y_1 + 0.50y_2 \tag{6.20}$$

Multiplying the first of Eqs. (6.15) by 3 and subtracting the second equation from it,

$$4x_1 = 3y_1 - y_2$$

or

$$x_1 = +0.75y_1 - 0.25y_2 \tag{6.21}$$

In matrix form, Eqs. (6.20) and (6.21) can be written as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} +0.75 & -0.25 \\ -0.50 & +0.50 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{6.22}$$

A more realistic set of equations would be

$$\begin{aligned} +4x_1 + 2x_2 \qquad\quad &= y_1 \\ +2x_1 + 10x_2 + 4x_3 &= y_2 \\ +\ 4x_2 + 5x_3 &= y_3 \end{aligned} \tag{6.23}$$

or

$$\begin{bmatrix} +4 & +2 & 0 \\ +2 & +10 & +4 \\ 0 & +4 & +5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{6.24}$$

The solution is approached by first defining the values of the determinant and the various cofactors:

$$D = +116$$

$$\begin{array}{lll} A_{11} = +34 & A_{12} = -10 & A_{13} = +8 \\ A_{21} = -10 & A_{22} = +20 & A_{23} = -16 \\ A_{31} = +8 & A_{32} = -16 & A_{33} = +36 \end{array} \tag{6.25}$$

For the problem in question, the inverse to three decimal places is the matrix

$$[A^{-1}] = \begin{bmatrix} +0.294 & -0.086 & +0.069 \\ -0.086 & +0.172 & -0.138 \\ +0.069 & -0.138 & +0.310 \end{bmatrix} \tag{6.26}$$

It should be noted that the original $[A]$ was symmetric. The inverse matrix $[A^{-1}]$ is also found to be symmetric. This is a basic property of square matrices: *Symmetrical matrices will always invert to symmetrical matrices.*

Since this method of inversion requires the use of Cramer's rule for evaluating simultaneous equations, and since it was demonstrated in Sec. 5.2 that the amount of work required to obtain a solution using that method can become prohibitive as the order of the equations increases, other methods of inverting matrices must be found.

6.3 *Elementary Row Operators*

METHOD #2

One of the simplest of the methods of obtaining the inverse of a matrix is that usually referred to as inversion by elementary row operators. Consider the set of equations

$$[A][X] = [I][Y] \tag{6.27}$$

Assume that both sides of each of these equations is altered by either or all of the following elementary operations: (1) premultiplication or division by a nonzero constant, or (2) the interchange of any two rows, or (3) the addition to any row of a multiple of any other row. Moreover, presuppose that these various elementary operations reduce the left-hand side of the original equation to the identity matrix $[I]$ times the column vector $[X]$. That is, for

$$\theta_1\theta_2 \cdots \theta_n[A][X] = \theta_1\theta_2 \cdots \theta_n[I][Y] \tag{6.28}$$

where $\theta_1, \theta_2, \ldots, \theta_n$ are the elementary row operators, if

$$\theta_1\theta_2 \cdots \theta_n[A] = [I] \tag{6.29}$$

then it is observed that

$$[I][X] = \theta_1\theta_2 \cdots \theta_n[I][Y] \tag{6.30}$$

or the inverse is

$$[A^{-1}] = \theta_1\theta_2 \cdots \theta_n[I] \tag{6.31}$$

It therefore can be stated that a sequence of elementary row operations on a matrix $[A]$ that transform it to the identity matrix $[I]$ will similarly transform an original identity matrix $[I]$ to the inverse $[A^{-1}]$.

To illustrate this, consider the set defined by Eqs. (6.24). The matrix in question is repeated as Eq. (6.32).

$$[A] = \begin{bmatrix} +4 & +2 & 0 \\ +2 & +10 & +4 \\ 0 & +4 & +5 \end{bmatrix} \tag{6.32}$$

The solution using elementary row operators is summarized in Table 6.1. The inverse of the matrix given as Eq. (6.32) is therefore

$$[A^{-1}] = \begin{bmatrix} +0.294 & -0.086 & +0.069 \\ -0.087 & +0.172 & -0.138 \\ +0.069 & -0.138 & +0.310 \end{bmatrix} \tag{6.33}$$

While certain other methods may require much less computational time, the elementary-row-operator method of inverting a matrix has much to offer. Not the least of these advantages is that it is easy to remember. Moreover, it is comparatively difficult to make a fundamental mistake when using this method.

6.4 Elimination Methods

SKIP

Chapter 5 was concerned with the solution of simultaneous equations which contained constants on the right-hand side. Therein, various elimination procedures were described. Extensive use of these methods will be made in this part of the discussion on matrix inversion.

It should be noted at the outset that the relative influence of any one of the "right-hand constants" can be determined by setting the value of that particular constant equal to 1, while all other elements in the column vector are equated to zero. Had the value of the constant in question been other than the assumed 1.0, say y_i, the corresponding x's would have been changed a proportional amount. The procedure for examining the influence of each of the other y_i positions would be the same. Therefore, to obtain the inverse of an $n \times n$ matrix by elimination methods, n separate simultaneous equations must be solved. In each of these a different position y_i is assumed equal to 1.0, and the others in that group are set equal to zero. Consider, for example, a 3×3 set. The three

MATRIX INVERSION BY ROW OPERATORS

Table 6.1

Line	Operation†	Original $[A]$			Original $[I]$			ΣH
(1)		+4	+2	0	1	0	0	+7.000
(2)		+2	+10	+4	0	1	0	+17.000
(3)		0	+4	+5	0	0	1	+10.000
(4)	(1) ÷ 4	+1.000	+0.500	0	+0.250	0	0	+1.750
(5)	(2) − 2 · (4)		+9.000	+4.000	−0.500	+1.000	0	+13.500
(6)	(3)		+4.000	+5.000	0	0	+1.000	+10.000
(7)	(5) ÷ 9.000		+1.000	+0.444	−0.056	+0.111	0	(499) +1.500
(8)	(6) − 4 · (7)			+3.224	+0.224	−0.444	+1.000	(4) +4.000
(9)	(8) ÷ 3.224			+1.000	+0.069	−0.138	+0.310	+1.241
Summary:								
(4)		+1.000	+0.500	0	+0.250	0	0	+1.750
(7)			+1.000	+0.444	−0.056	+0.111	0	+1.500
(9)				+1.000	+0.069	−0.138	+0.310	+1.241
(10)	2 · (4) − (7)	+2.000	0	−0.444	+0.556	−0.111	0	(1) +2.000
(11)	(10) + 0.444(9)	+2.000	0	0	+0.587	−0.172	+0.138	(3) +2.551
(12)	(11) ÷ 2.000	+1.000	0	0	+0.294	−0.086	+0.069	(7) +1.276
(13)	(7) − 0.444(9)		+1.000	0	−0.087	+0.172	−0.138	(7) +0.949
Summary:								
(12)		+1.000	0	0	+0.294	−0.086	+0.069	
(13)		0	+1.000	0	−0.087	+0.172	−0.138	
(9)		0	0	+1.000	+0.069	−0.138	+0.310	
		Final $[I]$			Final $[A^{-1}]$			

†Numbers in parentheses refer to the contents of the designated line.

simultaneous equations are

$$\begin{array}{lclcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1 & & = 0 & & = 0 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = 0 & \text{or} & = y_2 & \text{or} & = 0 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = 0 & & = 0 & & = y_3 \end{array} \qquad (6.34)$$

Possibly a more meaningful form of these equations is the following:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} y_1 \quad \text{or} \quad = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} y_2 \quad \text{or} \quad = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} y_3 \tag{6.35}$$

To indicate more formally the procedure being proposed, consider the following:

$$[A] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} y_1 + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} y_2 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} y_3 \tag{6.36}$$

Premultiplying each of the terms by $[A^{-1}]$,

$$\begin{aligned} [A^{-1}][A] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = [I] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= [A^{-1}] \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} y_1 + [A^{-1}] \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} y_2 \\ &\quad + [A^{-1}] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} y_3 \end{aligned} \tag{6.37}$$

It should be noted, however, that

$$[A^{-1}] \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & 0 & 0 \\ b_{31} & 0 & 0 \end{bmatrix} = \text{column vector corresponding to first column of inverse matrix} \tag{6.38}$$

$$[A^{-1}] \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & b_{12} & 0 \\ 0 & b_{22} & 0 \\ 0 & b_{32} & 0 \end{bmatrix} = \text{column vector corresponding to second column of inverse matrix} \tag{6.39}$$

$$[A^{-1}] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & b_{13} \\ 0 & 0 & b_{23} \\ 0 & 0 & b_{33} \end{bmatrix} = \text{column vector corresponding to third column of inverse matrix} \tag{6.40}$$

In other words, each of these column vectors is a column of $[A^{-1}]$ defined by the row in which the 1 appears. It should also be evident that each represents a solution vector of x's for the set of simultaneous linear equations whose column vector of constants is .

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

6.4.1 Gauss Elimination

Skip

To illustrate inversion by this method, consider the following matrix:

$$[A] = \begin{bmatrix} +3 & -3 & +5 \\ +1 & +2 & -6 \\ +2 & -1 & +3 \end{bmatrix} \tag{6.41}$$

The solution is carried out in Table 6.2, where three sets of constants have been assumed. Note the similarity with Table 6.1. The inverted matrix is

$$[A^{-1}] = \begin{bmatrix} 0 & +0.200 & +0.400 \\ -0.750 & -0.050 & +1.150 \\ -0.250 & -0.150 & +0.450 \end{bmatrix} \tag{6.42}$$

Table 6.2

	Coefficients of			Constants			
Line	x_1	x_2	x_3	y_1	y_2	y_3	ΣH
1	+3	−3	+5	1	0	0	
2	+1	+2	−6	0	1	0	
3	+2	−1	+3	0	0	1	
4	1	−1.000	+1.667	+0.333	0	0	
5		+3.000	−7.667	−0.333	+1.000	0	
6		+1.000	−0.334	−0.666	0	+1.000	
7		1	−2.556	−0.111	+0.333	0	
8			+2.222	−0.555	−0.333	+1.000	
9			1	−0.250	−0.150	+0.450	
10		1		−0.750	−0.050	+1.150	
11	1			0	+0.200	+0.400	

To verify that this is in fact the inverse, the matrix product of $[A]$ and $[A^{-1}]$ should equal the identity matrix $[I]$.

$$\begin{bmatrix} +3 & -3 & +5 \\ +1 & +2 & -6 \\ +2 & -1 & +3 \end{bmatrix} \begin{bmatrix} 0 & +0.200 & +0.400 \\ -0.750 & -0.050 & +1.150 \\ -0.250 & -0.150 & +0.450 \end{bmatrix}$$

$$= \begin{bmatrix} +1.000 & 0 & 0 \\ 0 & +1.000 & 0 \\ 0 & 0 & +1.000 \end{bmatrix} \tag{6.43}$$

Within the accuracy of the calculations, the identity matrix is realized.

While conceptually the solution of n separate simultaneous equations is required to invert an $n \times n$ matrix by elimination methods, it should be evident from the above that this does not mean that there will be n times the amount of work required. The entire left-hand part of the tabular solution will apply to each of the sets of solution.

6.4.2 Cholesky's Method — SKIP

Table 6.3 summarizes the matrix inversion of Eq. (6.41) using Cholesky's method. (Note the similarity with Table 5.5.) The upper unit triangular set of equations is, then,

Table 6.3

i \ j	1	2	3	Constants y_1	Constants y_2	Constants y_3	1	2	3	Constants 4	Constants 5	Constants 6
1	+3	−3	+5	1	0	0	+3.00	−1.00	+1.67	+0.33	0	0
2	+1	+2	−6	0	1	0	+1.00	+3.00	−2.56	−0.11	+0.33	0
3	+2	−1	+3	0	0	1	+2.00	+1.00	+2.23	−0.25	−0.15	+0.45

|←a_{ij}'s→|←c's→|←$(\ell_{ij}\backslash t_{ij})$'s→|←$k$'s→|

$$\begin{array}{r} x_1 - 1.00x_2 + 1.67x_3 \\ x_2 - 2.56x_3 \\ x_3 \end{array}\ \overset{y_1}{\boxed{\begin{array}{l} = +0.33 \\ = -0.11 \\ = -0.25 \end{array}}}\ \text{or}\ \overset{y_2}{\boxed{\begin{array}{l} = \quad 0 \\ = +0.33 \\ = -0.15 \end{array}}}\ \text{or}\ \overset{y_3}{\boxed{\begin{array}{l} = \quad 0 \\ = \quad 0 \\ = +0.45 \end{array}}} \tag{6.44}$$

The corresponding equations for the x's in terms of the y's are given as

$$\begin{aligned} x_1 &= 0y_1 + 0.20y_2 + 0.40y_3 \\ x_2 &= -0.75y_1 - 0.05y_2 + 1.15y_3 \\ x_3 &= -0.25y_1 - 0.15y_2 + 0.45y_3 \end{aligned} \tag{6.45}$$

The inverse is, then,

$$[A^{-1}] = \begin{bmatrix} 0 & +0.20 & +0.40 \\ -0.75 & -0.05 & +1.15 \\ -0.25 & -0.15 & +0.45 \end{bmatrix} \tag{6.46}$$

Again, it should be recognized that solution for the n separate sets of constants does not require n times the amount of work. The computational matrices $[L]$ and $[T]$ will be the same for all sets of constants.

SKIP

6.5 *Inversion of Matrices by Iteration or Relaxation*

Iteration methods, like relaxation methods, require independent sets of calculations for each assumed set of "right-hand-side constants." Therefore, in general, n times the amount of work required to obtain one set of solutions (i.e., one column of the inverse) will be required to invert an $n \times n$ matrix. It should be noted, however, that should a powerful group-relaxation-operation table be found, the labor involved in relaxation may not be prohibitive.

6.6 *Matrix Inversion by Partitioning*

READ METHOD #3

It was pointed out in Sec. 4.5 that certain matrices can be naturally partitioned. For such cases inversion may be accomplished by manipulation of these smaller submatrices rather than dealing with the total aggregation.

Assume that a given $n \times n$ matrix can be naturally partitioned so that the leading-diagonal subsets are square.

$$[A] = \left[\begin{array}{c:c} [A_{11}] & [A_{12}] \\ \hdashline [A_{21}] & [A_{22}] \end{array}\right] \tag{6.47}$$

The orders of $[A_{11}]$ and $[A_{22}]$ are presumed to be p and q, such that $p + q = n$. The inverse matrix is also presumed to be partitionable, as indicated in Eq. (6.48).

$$[A^{-1}] = \left[\begin{array}{c:c} [B_{11}] & [B_{12}] \\ \hdashline [B_{21}] & [B_{22}] \end{array}\right] \tag{6.48}$$

Here, it is further specified that $[B_{11}]$ and $[B_{22}]$ are also, respectively, of order p and q.

From Eq. (6.9), the condition for matrix inversion is

$$[A][A^{-1}] = [I] \tag{6.49}$$

Substituting the values defined by Eqs. (6.47) and (6.48) in Eq. (6.49) gives

$$\left[\begin{array}{c:c} [A_{11}] & [A_{12}] \\ \hdashline [A_{21}] & [A_{22}] \end{array}\right] \left[\begin{array}{c:c} [B_{11}] & [B_{12}] \\ \hdashline [B_{21}] & [B_{22}] \end{array}\right] = \left[\begin{array}{c:c} [I] & [0] \\ \hdashline [0] & [I] \end{array}\right] \tag{6.50}$$

Multiplying the two matrices in the left-hand member and equating the result to the right-hand member yields

$$[A_{11}][B_{11}] + [A_{12}][B_{21}] = [I] \tag{6.51}$$

$$[A_{11}][B_{12}] + [A_{12}][B_{22}] = [0] \tag{6.52}$$

$$[A_{21}][B_{11}] + [A_{22}][B_{21}] = [0] \tag{6.53}$$

$$[A_{21}][B_{12}] + [A_{22}][B_{22}] = [I] \tag{6.54}$$

Multiplying Eq. (6.53) by $[A_{12}][A_{22}^{-1}]$,

$$[A_{12}][A_{22}^{-1}][A_{21}][B_{11}] + [A_{12}][A_{22}^{-1}][A_{22}][B_{21}] = 0 \tag{6.55}$$

Noting that $[A_{22}^{-1}][A_{22}] = [I]$,

$$[A_{12}][A_{22}^{-1}][A_{21}][B_{11}] + [A_{12}][B_{21}] = 0 \tag{6.56}$$

Subtracting Eq. (6.56) from (6.51) yields

$$[A_{11}][B_{11}] - [A_{12}][A_{22}^{-1}][A_{21}][B_{11}] = [I]$$

or

$$[[A_{11}] - [A_{12}][A_{22}^{-1}][A_{21}]][B_{11}] = [I] \tag{6.57}$$

Solving Eq. (6.57) for $[B_{11}]$, and noting that in general if $[M][Z] = [I]$, then $[Z] = [M^{-1}] = [M]^{-1}$, give

NEED INVERSES OF A_{11} & A_{2}

From Eq. (6.53),

$$\boxed{\begin{aligned} [B_{11}] &= [[A_{11}] - [A_{12}][A_{22}^{-1}][A_{21}]]^{-1}\ \uparrow\text{NOTE} \\ [B_{21}] &= [-[A_{22}^{-1}][A_{21}][B_{11}]] \end{aligned}} \qquad \begin{aligned} &(6.58) \\ &(6.59) \end{aligned}$$

In a similar manner, Eqs. (6.52) and (6.54) will yield Eqs. (6.60) and (6.61).

and

$$\boxed{\begin{aligned} [B_{22}] &= [[A_{22}] - [A_{21}][A_{11}^{-1}][A_{12}]]^{-1}\ \uparrow\text{NOTE} \\ [B_{12}] &= [-[A_{11}^{-1}][A_{12}][B_{22}]] \end{aligned}} \qquad \begin{aligned} &(6.60) \\ &(6.61) \end{aligned}$$

The inversion of an $n \times n$ matrix by partitioning thus reduces to the inversion of four matrices. Two of these will be of order p. The other two will be of order q. Several matrix multiplications and additions are also required.

To illustrate inversion by this method, consider the matrix defined by

$$[A] = \left[\begin{array}{ccc|cc} +1 & 0 & +1 & 0 & 0 \\ 0 & +2 & 0 & +1 & 0 \\ -1 & 0 & +1 & 0 & +1 \\ \hline 0 & 0 & 0 & +2 & 0 \\ 0 & 0 & 0 & 0 & +2 \end{array}\right] = \left[\begin{array}{c|c} [A_{11}] & [A_{12}] \\ \hline [A_{21}] & [A_{22}] \end{array}\right] \tag{6.62}$$

The required inversions of the original submatrices are as given in Eqs. (6.63).

$$[A_{11}^{-1}] = \begin{bmatrix} +0.50 & 0 & -0.50 \\ 0 & +0.50 & 0 \\ +0.50 & 0 & +0.50 \end{bmatrix} \qquad [A_{22}^{-1}] = \begin{bmatrix} +0.50 & 0 \\ 0 & +0.50 \end{bmatrix} \tag{6.63}$$

Considering Eq. (6.58),

$$[A_{12}][A_{22}^{-1}][A_{21}] = \begin{bmatrix} 0 & 0 \\ +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} +0.50 & 0 \\ 0 & +0.50 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6.64}$$

Therefore,

$$[B_{11}] = [A_{11}^{-1}] = \begin{bmatrix} +0.50 & 0 & -0.50 \\ 0 & +0.50 & 0 \\ +0.50 & 0 & +0.50 \end{bmatrix} \tag{6.65}$$

For Eq. (6.59),

$$[A_{22}^{-1}][A_{21}][B_{11}] = \begin{bmatrix} +0.50 & 0 \\ 0 & +0.50 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} +0.50 & 0 & -0.50 \\ 0 & +0.50 & 0 \\ +0.50 & 0 & +0.50 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6.66}$$

or

$$[B_{21}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6.67}$$

Consider Eq. (6.60).

$$[A_{21}][A_{11}^{-1}][A_{12}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} +0.50 & 0 & -0.50 \\ 0 & +0.50 & 0 \\ +0.50 & 0 & +0.50 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ +1 & 0 \\ 0 & +1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{6.68}$$

or

$$[B_{22}] = [A_{22}^{-1}] = \begin{bmatrix} +0.50 & 0 \\ 0 & +0.50 \end{bmatrix} \tag{6.69}$$

Equation (6.61) requires that

$$[A_{11}^{-1}][A_{12}][B_{22}] = \begin{bmatrix} +0.50 & 0 & -0.50 \\ 0 & +0.50 & 0 \\ +0.50 & 0 & +0.50 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} +0.50 & 0 \\ 0 & +0.50 \end{bmatrix} = \begin{bmatrix} 0 & -0.25 \\ +0.25 & 0 \\ 0 & +0.25 \end{bmatrix} \tag{6.70}$$

or

$$[B_{12}] = \begin{bmatrix} 0 & +0.25 \\ -0.25 & 0 \\ 0 & -0.25 \end{bmatrix} \tag{6.71}$$

The inverted matrix is, therefore,

$$[A^{-1}] = \left[\begin{array}{ccc:cc} +0.50 & 0 & -0.50 & 0 & +0.25 \\ 0 & +0.50 & 0 & -0.25 & 0 \\ +0.50 & 0 & +0.50 & 0 & -0.25 \\ \hdashline 0 & 0 & 0 & +0.50 & 0 \\ 0 & 0 & 0 & 0 & +0.50 \end{array}\right] \tag{6.72}$$

To verify that this is in fact the inverse, the matrix product of $[A]$ and $[A^{-1}]$ is shown in Eq. (6.73).

$$\begin{bmatrix} +1 & 0 & +1 & 0 & 0 \\ 0 & +2 & 0 & +1 & 0 \\ -1 & 0 & +1 & 0 & +1 \\ 0 & 0 & 0 & +2 & 0 \\ 0 & 0 & 0 & 0 & +2 \end{bmatrix} \begin{bmatrix} +0.50 & 0 & -0.50 & 0 & +0.25 \\ 0 & +0.50 & 0 & -0.25 & 0 \\ +0.50 & 0 & +0.50 & 0 & -0.25 \\ 0 & 0 & 0 & +0.50 & 0 \\ 0 & 0 & 0 & 0 & +0.50 \end{bmatrix}$$

$$= \begin{bmatrix} +1 & 0 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 & 0 \\ 0 & 0 & +1 & 0 & 0 \\ 0 & 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & +1 \end{bmatrix} \tag{6.73}$$

Prob. 6.1: By Cramer's rule, elementary row operators, Gauss elimination, and Cholesky's method, invert the following matrices:

(*a*) $\begin{bmatrix} +1 & 0 & -1 \\ 0 & +1 & 0 \\ +1 & 0 & +1 \end{bmatrix}$ (*b*) $\begin{bmatrix} +1 & +3 & +2 \\ 0 & +1 & +3 \\ 0 & 0 & +1 \end{bmatrix}$

(*c*) $\begin{bmatrix} +0.25 & +0.30 & -0.10 \\ +0.20 & -0.50 & +0.15 \\ -0.15 & +0.35 & +0.60 \end{bmatrix}$ (*d*) $\begin{bmatrix} -0.294 & +0.086 & -0.069 \\ +0.086 & -0.172 & +1.138 \\ -0.069 & +0.138 & -0.310 \end{bmatrix}$

Prob. 6.2: For the following 3 × 3 set of equations relating x's and y's, determine by Cholesky's method the y's as explicit functions of the x's.

$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 1y_1 + 2y_2 \\ 2x_1 + 5x_2 - 2x_3 &= 2y_1 + 3y_2 + 1y_3 \\ 1x_1 - 2x_2 + 4x_3 &= + 1y_2 + 5y_3 \end{aligned}$$

Prob. 6.3: By partitioning, determine the inverse of the following matrices:

(*a*)

$$[A] = \begin{bmatrix} 1 & 0 & 3 & 0 \\ 2 & 2 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

(*b*)

$$[A] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{array}\right]$$

Prob. 6.4: Given the symmetrical 4 × 4 tridiagonal system shown, determine ~~the~~ inverse of the set by partitioning.

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix}$$

Prob. 6.5: It is assumed that a matrix can be partitioned so that $[A_{12}] = 0$. In terms of the submatrices $[A_{11}]$, $[A_{21}]$, and $[A_{22}]$, what will be the elements of the inverse of such a system? Using these results, determine the inverse of the following matrix:

$$\left[\begin{array}{ccc:cc} 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \hdashline 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 2 \end{array}\right]$$

Prob. 6.6: If in Prob. 6.5 the matrix had been "bordered" [see Eq. (4.56)], as shown in the following set,

$$\left[\begin{array}{cccc:c} 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hdashline 0 & 1 & 0 & 0 & 2 \end{array}\right]$$

would the derived equations also apply, and if so, would any computational advantage be realized?

Prob. 6.7: It is assumed that an arbitrary lower triangular matrix will invert to another lower triangular matrix. For a 3 × 3 system, develop the equation defining the elements of the inverse. Are there any special limitations on this set of equations?

Prob. 6.8: In Example 4.4, it was presumed that voltages E_1 and E_2 were specified and that values of the unknown currents I_1, I_2, I_3, and I_4 were desired. Assuming that all the resistances are equal, i.e.,

$$R_1 = R_2 = R_3 = R_4 = R_5 = R_6 = R$$

and that the currents are known, determine the expressions for E_1 and E_2 in terms of the I's and R.

Prob. 6.9: Example 4.8 was concerned with the transformation of a generalized force-couple system in rectangular coordinates x, y, and z to a parallel axis system $\bar{x}$, $\bar{y}$, and $\bar{z}$. The equations defining the transformation are given as (4.101). Obtain the inverse of this system; i.e., define

V_i as an explicit function of $\bar{V}_i$. From geometrical considerations and Eqs. (4.101), verify these new equations.

6.7 *Computer Solutions*

More often than not, engineering and other problems in applied science are "directionally formulated"; that is, one of the quantities in question is more easily defined in terms of the remaining variables than vice versa. For example, in structural analysis, deflections are readily described in terms of geometrical properties of the various members and the imposed loading. In electric circuits, voltages can easily be defined in terms of resistances and currents. While this is true, the reverse (or inverse) formulation is frequently the more desirable. This is especially so when design problems are considered.

Of the several methods for the inversion of matrices that have been discussed, perhaps the most adaptable to computer solutions are those based on elimination methods. The case that has been chosen for illustration is that of repeated application of Cholesky's method.

It will be recalled from Sec. 6.4.2 and Table 6.3 that Cholesky's method, extended to the inverse problem, requires that the coefficient matrix $[A]$ be augmented by an identity matrix $[I]$ of order n. By operating on each column of this identity set as if it were a separate column vector of constants, that is, by considering n different sets of simultaneous equations, the n different columns of the inverse $[A^{-1}]$ can be established.

It is necessary, in writing the computer program, to ensure that each specific column of the identity matrix in the augmented set corresponds to its respective column in the inverse. To facilitate this, a special counter denoted as m will be used to define the columns of the inverse matrix.

A flow diagram that will invert a matrix by use of Cholesky's method is shown in Fig. 6.1. It will be noted that use is made of the routine for the solution of simultaneous equations that was presented earlier as Fig. 5.18. It has been assumed that a simultaneous-equation program is available for use as a subroutine.

The notes in Fig. 6.1 refer to the following:

1. As stated above, m is a counter which defines the column of the inverse matrix currently being calculated. It will vary from 1 to n.
2. At this stage, beginning with $i = 1$, a test is required to determine whether $i = m$. If such is found to be the case, the corresponding element in $[C]$ or c_i must be set equal to 1. All other c_i's must equal zero. In this manner the required column matrices c_i are generated.

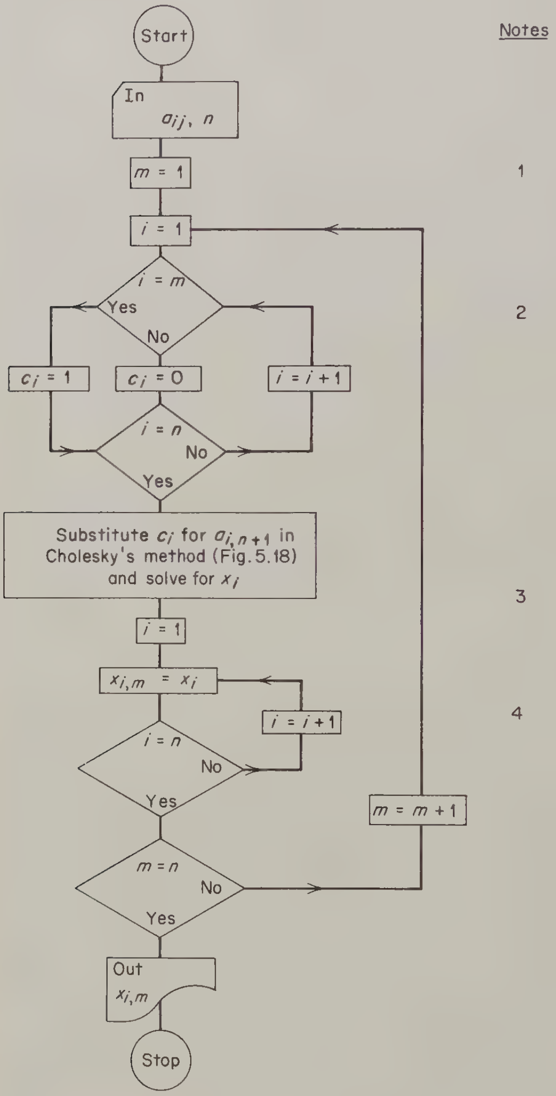


Figure 6.1 Flow diagram for matrix inversion—Cholesky's method.

3. The actual flow diagram for the solution of the simultaneous equations is presumed known. (Entry into that flow diagram is made at note 2 in Fig. 5.11.) Exit from the simultaneous-equation routine to this program would be just prior to the stage where the x_i are to be put out (Fig. 5.16).
4. An $n \times n$ matrix is formed, the columns of which are the values of

the x_i's just determined. When all n columns have been established, that is, when $m = n$, the values of $x_{i,m}$ are read out as the elements of the inverse $[A^{-1}]$.

Prob. 6.10: Construct a flow diagram for the inversion, by partitioning, of an nth-order set using Eqs. (6.58) to (6.61). Introduce checks to ensure that the matrix operations can be realized.

eigenvalue problems

7

7.1 Introduction

In all the preceding work on matrices and simultaneous equations, it has been assumed that the coefficient matrix is *nonsingular.* That is, the value of the determinant of the coefficients is other than zero. It should also have been noted that in each case the elements of the various assumed matrices were constants. Another observation that could have been made is that the equations considered were nonhomogeneous.

Eigenvalue problems are concerned specifically with the solution of sets of homogeneous equations (i.e., all elements of the column vector of constants equal zero). For illustration of the general problem, consider the set of equations

$$\begin{aligned} p_{11}x_1 + p_{12}x_2 + \cdots + p_{1n}x_n &= 0 \\ p_{21}x_1 + p_{22}x_2 + \cdots + p_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots&\cdots \\ p_{n1}x_1 + p_{n2}x_2 + \cdots + p_{nn}x_n &= 0 \end{aligned} \tag{7.1}$$

It will not be assumed necessarily that any or all of the coefficient elements p_{ij} are constants.

One obvious, though trivial, set of solutions to the simultaneous equations (7.1) is

$$x_1 = x_2 = x_3 \cdots = x_n = 0 \tag{7.2}$$

There are, however, other possibilities. From Eq. (5.17) (Cramer's rule), x_j can be defined as

$$x_j = \frac{D_j}{D} \tag{7.3}$$

where D is the value of the determinant of the coefficient matrix $[p_{ij}]$ and D_j is the value of the same determinant with the jth column replaced by the column of zeros (the vector of constants). Evaluating D_j by expansion about the jth column (now replaced by zeros),

$$D_j = 0 \tag{7.4}$$

Therefore, for x_j to have a value other than zero, a necessary and sufficient condition for solution is that the determinant of the coefficient matrix equal zero.

$$D = 0 \tag{7.5}$$

or

$$D = |P| = \begin{vmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{vmatrix} = 0 \tag{7.6}$$

It will now be assumed that the coefficient matrix $[P]$ is actually the difference between two other matrices $[A]$ and $[\underline{B}]$. Moreover, it will further be presumed that $[\underline{B}]$ is a coefficient matrix $[B]$ multiplied by an unknown scalar λ. $[P]$, then, is defined as

$$[P] = [A] - \lambda[B] \tag{7.7}$$

or

$$p_{ij} = a_{ij} - \lambda b_{ij} \qquad (i, j = 1, 2, \ldots, n) \tag{7.8}$$

and the set of simultaneous equations under consideration is given as (7.9) or (7.10).

$$[[A] - \lambda[B]][X] = 0 \tag{7.9}$$

$$\begin{aligned} (a_{11} - \lambda b_{11})x_1 + (a_{12} - \lambda b_{12})x_2 + \cdots + (a_{1n} - \lambda b_{1n})x_n &= 0 \\ (a_{21} - \lambda b_{21})x_1 + (a_{22} - \lambda b_{22})x_2 + \cdots + (a_{2n} - \lambda b_{2n})x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ (a_{n1} - \lambda b_{n1})x_1 + (a_{n2} - \lambda b_{n2})x_2 + \cdots + (a_{nn} - \lambda b_{nn})x_n &= 0 \end{aligned} \tag{7.10}$$

Here again the necessary condition for a solution is that the determinant of the coefficient matrix equal zero. That is,

$$|A - \lambda B| = 0 \tag{7.11}$$

or

$$\begin{vmatrix} a_{11} - \lambda b_{11} & a_{12} - \lambda b_{12} & \cdots & a_{1n} - \lambda b_{1n} \\ a_{21} - \lambda b_{21} & a_{22} - \lambda b_{22} & \cdots & a_{2n} - \lambda b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} - \lambda b_{n1} & a_{n2} - \lambda b_{n2} & \cdots & a_{nn} - \lambda b_{nn} \end{vmatrix} = 0 \tag{7.12}$$

By expanding Eqs. (7.11) or (7.12), there is obtained an nth-degree polynomial in the unknown λ's. This can be written in the form

$$\lambda^n + q_1\lambda^{n-1} + q_2\lambda^{n-2} + \cdots + q_{n-1}\lambda^1 + q_n = 0 \tag{7.13}$$

The n roots of this polynomial are known as the eigenvalues. They correspond to those particular λ values which allow the determinant of the coefficient matrix to equal zero. It should be noted, however, that these eigenvalues may not be all real or distinct.

Corresponding to each of the n eigenvalues is an *eigenvector*, that is, the column vector of x values. It should be noted that since a solution has essentially been forced (λ chosen so that $D = 0$), no one unique set of x's can be found. The ratio between the various x values can, however, be established, as will be proved in the following paragraph.

From Eq. (7.9),

$$[A][X] = \lambda_1[B][X] \tag{7.14}$$

which assumes that $[X]$ is an eigenvector corresponding to the eigenvalue λ_1. Assume now that $[aX]$ is also an eigenvector, where a is a nonzero scalar.

$$[A][aX] = a[A][X] = a\lambda_1[B][X] = \lambda_1[B][aX] \tag{7.15}$$

That is, eigenvectors can be determined only to a proportionality.

In many real problems the coefficient matrix $[B]$ is the identity matrix $[I]$. When such is the case, the resulting polynomial in λ is often referred to as the *characteristic equation* or *secular equation.* The corresponding eigenvalues are frequently denoted as *proper values.*

Thus far it has been presumed that the λ values are scalar multipliers of the coefficient matrix $[B]$. This need not be the case. λ may be contained directly in matrix $[B]$ as the argument of a trigonometric, logarithmic, Bessel, or any other function. Consider for illustration the following equations, which define a rather elementary buckling situation (see Example 4.1).

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ \lambda & 0 & 1 & 0 \\ \sin\lambda & \cos\lambda & 1 & 1 \\ \sin\lambda & \cos\lambda & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7.16}$$

The two matrices $[A]$ and $[B]$ are

$$[A] = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.17}$$

$$[B] = -1 \begin{bmatrix} 0 & 0 & 0 & 0 \\ \lambda & 0 & 0 & 0 \\ \sin\lambda & \cos\lambda & 0 & 0 \\ \sin\lambda & \cos\lambda & 0 & 0 \end{bmatrix} \tag{7.18}$$

where

$$[A - B][X] = [0] \tag{7.19}$$

7.2 *Examples of Eigenvalue Problems*

Many engineering problems require the determination of eigenvalues and their corresponding eigenvectors. Three very general areas in which this becomes the fundamental consideration are as follows: (1) natural frequencies of vibrations, (2) buckling, and (3) resonance in acoustical and electrical circuits. A wide range of physical situations would be encompassed in each of these. For example, consider just a few of those associated with the first area: (1) the natural vibration of systems of masses and springs, (2) the flutter of airplane wings, (3) the vibration of membranes, (4) the oscillation of suspension bridges, (5) the torsional vibration of multicylinder engines, and (6) structural response to earthquakes.

Example 7.1 As a first example, consider the set of two simultaneous equations

$$\begin{aligned} x_1 + 2x_2 &= \lambda x_1 \\ 2x_1 + x_2 &= \lambda x_2 \end{aligned} \qquad \text{or} \qquad \begin{bmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{7.20}$$

Since these equations are homogeneous, the necessary condition for a solution is that the value of the determinant of the coefficient matrix be zero:

$$\begin{vmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} = 0 \tag{7.21}$$

Expanding Eq. (7.21) gives the characteristic polynomial equation as

$$\lambda^2 - 2\lambda - 3 = 0 \qquad \text{or} \qquad (\lambda + 1)(\lambda - 3) = 0 \tag{7.22}$$

The eigenvalues are therefore

$$\lambda_1 = +3 \qquad \text{and} \qquad \lambda_2 = -1 \tag{7.23}$$

To determine the eigenvectors for these roots (eigenvalues) it is necessary to assume the value of one of the x's; for example, select $x_1 = +1.0$ for $\lambda_1 = +3$. From the first of equations (7.20), the x_2 value is found to be

$$x_2 = +1.0 \tag{7.24}$$

As a check, the same values of $x_1 = 1.0$ and $\lambda_1 = +3$ can be substituted in the second of the given equations. Here, again, it is found that $x_2 = +1.0$. Therefore the first eigenvalue and corresponding eigenvector are

$$\lambda_1 = +3 \qquad [X] = \begin{bmatrix} +1.0 \\ +1.0 \end{bmatrix} \tag{7.25}$$

For the negative root $\lambda_2 = -1$, the second eigenvalue, the solution is

$$\lambda_2 = -1 \qquad [X] = \begin{bmatrix} +1.0 \\ -1.0 \end{bmatrix} \tag{7.26}$$

Example 7.2

$$\begin{bmatrix} -1-\lambda & +2 & +2 \\ 0 & -2-\lambda & -1 \\ -4 & -2 & +3-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{7.27}$$

Equating the determinant of the coefficient matrix to zero gives

$$(-1-\lambda)\begin{vmatrix} -2-\lambda & -1 \\ -2 & +3-\lambda \end{vmatrix} - 4\begin{vmatrix} +2 & +2 \\ -2-\lambda & -1 \end{vmatrix} = 0 \tag{7.28}$$

or

$$\lambda^3 - \lambda = 0 = \lambda(\lambda - 1)(\lambda + 1) \tag{7.29}$$

The eigenvalues are therefore $\lambda_1 = 0$, $\lambda_2 = +1.0$, and $\lambda_3 = -1.0$. The eigenvectors can be determined in the same manner as in Example 7.1. The solutions are summarized in Eqs. (7.30) to (7.32):

$$\lambda_1 = 0 \qquad [X] = \begin{bmatrix} +1.0 \\ -0.5 \\ +1.0 \end{bmatrix} \tag{7.30}$$

$$\lambda_2 = +1.0 \qquad [X] = \begin{bmatrix} +1.0 \\ -0.5 \\ +1.5 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} +0.67 \\ -0.33 \\ +1.00 \end{bmatrix} \tag{7.31}$$

$$\lambda_3 = -1.0 \qquad [X] = \begin{bmatrix} -1.5 \\ +1.0 \\ -1.0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} +1.00 \\ -0.67 \\ +0.67 \end{bmatrix} \tag{7.32}$$

It should be noted that two eigenvectors have been listed for each of the last two eigenvalues. The first of these corresponds to the assumed +1.0 as one of the component values (x_1 in the second and x_2 in the third). The second sets of eigenvectors were obtained by normalizing with respect to the largest component.

Example 7.3 The problem under consideration is the determination of the natural frequencies of vibration (i.e., eigenvalues) of the two-mass

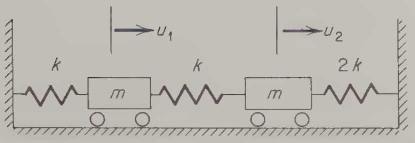


Figure 7.1

spring-connected system shown in Fig. 7.1. The corresponding eigenvectors are also desired. If a disturbance from the neutral position is assumed, the free-body diagrams for each of the masses (including d'Alembert forces) are those shown in Fig. 7.2. It has been presumed that $u_2 > u_1$ and that friction can be neglected.

From equilibrium of horizontal forces, the equations given as (7.33) are obtained.

$$\begin{aligned} -ku_1 + k(u_2 - u_1) &= m\ddot{u}_1 \\ -k(u_2 - u_1) - 2k(u_2) &= m\ddot{u}_2 \end{aligned} \tag{7.33}$$

For the natural frequencies, the displacement quantities (u_1 and u_2) can be related to the amplitudes of vibration (a_1 and a_2), the natural (or circular) frequencies (ω), and a phase angle (ϕ) according to Eqs. (7.34).

$$\begin{aligned} u_1 &= a_1 \sin(\omega t + \phi) \\ u_2 &= a_2 \sin(\omega t + \phi) \end{aligned} \tag{7.34}$$

Substitution of Eqs. (7.34) in Eqs. (7.33) yields

$$-ka_1 \sin(\omega t + \phi) + k(a_2 - a_1) \sin(\omega t + \phi) = -ma_1\omega^2 \sin(\omega t + \phi)$$

$$-k(a_2 - a_1) \sin(\omega t + \phi) - 2ka_2 \sin(\omega t + \phi) = -ma_2\omega^2 \sin(\omega t + \phi)$$

or

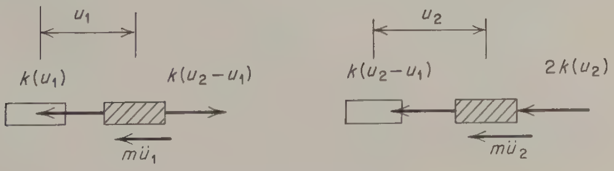
$$\begin{aligned} a_1(-2 + \lambda) + a_2(1) &= 0 \\ a_1(1) + a_2(-3 + \lambda) &= 0 \end{aligned} \tag{7.35}$$

where

$$\lambda = \frac{m\omega^2}{k} \tag{7.36}$$

The necessary (and sufficient) condition for Eqs. (7.35) to have other than a trivial solution is that the value of the determinant of the coeffi-

Figure 7.2



cient matrix be zero. That is,

$$\begin{vmatrix} -2+\lambda & 1 \\ 1 & -3+\lambda \end{vmatrix} = \begin{vmatrix} 2-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = 0 \tag{7.37}$$

The characteristic polynomial equation is therefore

$$\lambda^2 - 5\lambda + 5 = 0 \tag{7.38}$$

The eigenvalues, which would be the roots of Eq. (7.38), are

$$\begin{aligned} \lambda_1 &= +\frac{5}{2} - \frac{\sqrt{5}}{2} = +1.382 \\ \lambda_2 &= +\frac{5}{2} + \frac{\sqrt{5}}{2} = +3.618 \end{aligned} \tag{7.39}$$

The solution can, therefore, be summarized as in (7.40).

$$\begin{array}{lcl} \lambda_1 = +1.382 & & \lambda_2 = +3.618 \\ \omega_1 = \sqrt{1.382\dfrac{k}{m}} & & \omega_2 = \sqrt{3.618\dfrac{k}{m}} \\ [X] = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} +1.000 \\ +0.618 \end{bmatrix} & \text{and} & [X] = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.618 \\ +1.000 \end{bmatrix} \end{array} \tag{7.40}$$

A qualitative picture of what these mean physically is shown in Fig. 7.3.

Example 7.4 A flagpole type of uniform cross-section column is subjected to a concentrated vertical force at its uppermost point, as shown in Fig. 7.4. It is assumed that the vertical force will remain vertical. The differential equation governing the solution of this problem is

$$\frac{d^4y}{dx^4} + k^2\frac{d^2y}{dx^2} = 0 \tag{7.41}$$

where

$$k^2 = \frac{P}{EI} \tag{7.42}$$

The solution to this homogeneous differential equation is

$$y = C_1 \sin kx + C_2 \cos kx + C_3\frac{x}{L} + C_4 \tag{7.43}$$

The boundary conditions that must be satisfied are the following:

$$\begin{aligned} &\text{At } x = 0: \quad y = 0 \\ &\qquad (0)C_1 + (1)C_2 + (0)C_3 + (1)C_4 = 0 \\ &\text{At } x = 0: \quad y' = 0 \\ &\qquad (k)C_1 + (0)C_2 + \left(\frac{1}{L}\right)C_3 + (0)C_4 = 0 \\ &\text{At } x = L: \quad y'' = 0 \\ &\qquad (-k^2 \sin kL)C_1 + (-k^2 \cos kL)C_2 + (0)C_3 + (0)C_4 = 0 \end{aligned} \tag{7.44}$$

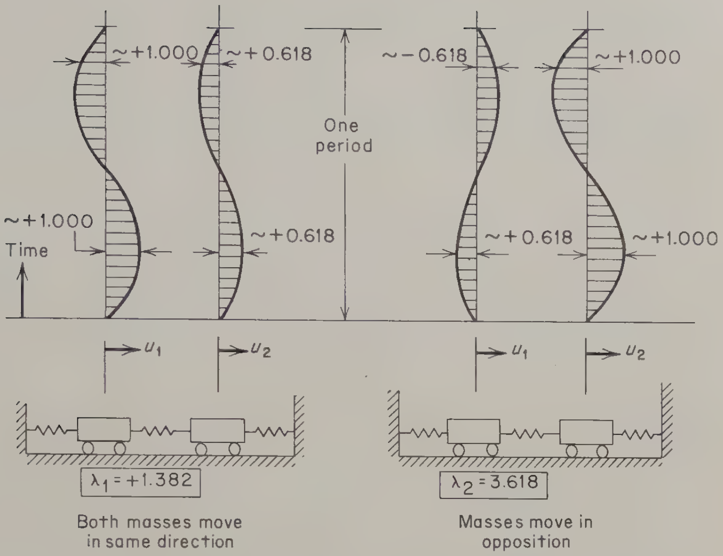


Figure 7.3

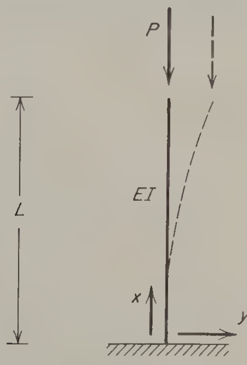
At $x = L$: $y''' + k^2y' = 0$

$$(0)C_1 \qquad + \qquad (0)C_2 \qquad + \left(\frac{k_2}{L}\right) C_3 + (0)C_4 = 0$$

By defining $kL = \lambda$, Eqs. (7.44) can be written as

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ \lambda & 0 & 1 & 0 \\ \sin\lambda & \cos\lambda & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7.45}$$

Figure 7.4



For a nontrivial solution to exist, the determinant of the coefficient matrix must equal zero. This gives, on expansion,

$$\lambda \cos \lambda = 0 \tag{7.46}$$

But $\lambda = 0$ will also result in a trivial solution (that is, $P = 0$). Therefore,

$$\cos \lambda = 0 \tag{7.47}$$

or

$$\lambda = (2n + 1)\frac{\pi}{2} \qquad (n = 0, 1, \ldots) \tag{7.48}$$

The smallest eigenvalue (that is, the first possibility of buckling) occurs for the case where $n = 0$.

$$\lambda_1 = \frac{\pi}{2} = k_1 L = \sqrt{\frac{P_1}{EI}}\,L \tag{7.49}$$

or

$$P_1 = \frac{\pi^2 EI}{4L^2} \tag{7.49a}$$

From the third of Eqs. (7.45), if it is noted from Eq. (7.47) that $\cos \lambda = 0$, $C_1 = 0$. From the fourth of these equations, $C_3 = 0$. Therefore, from the first equation,

$$C_2 = -C_4 \tag{7.50}$$

or the deflection equation is

$$y = C_4\left[1 - \cos \lambda \left(\frac{x}{L}\right)\right] \tag{7.51}$$

For the case $\lambda_1 = 0.5\pi$, the deflection equation is

$$y = C_4\left[1 - \cos \frac{\pi}{2}\left(\frac{x}{L}\right)\right] \tag{7.52}$$

The next larger eigenvalue corresponds to the case where n [in Eq. (7.48)] equals 1. That is,

$$\lambda_2 = \frac{3}{2}\pi \qquad \text{or} \qquad P_2 = \frac{9\pi^2 EI}{4L^2} \tag{7.52a}$$

The corresponding deflection equation is

$$y = C_4\left[1 - \cos \frac{3\pi}{2}\left(\frac{x}{L}\right)\right] \tag{7.52b}$$

The deflection configurations corresponding to these buckling solutions [Eqs. (7.52) and (7.52*b*)] are shown in Fig. 7.5. Here again, it should be noted that only relative values are known.

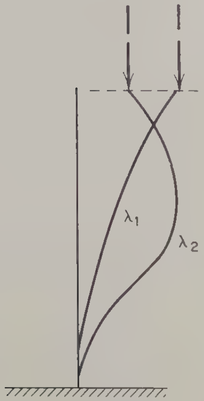


Figure 7.5

7.3 *Determination of Eigenvalues by Iteration*

Frequently, engineering problems require the determination of either the greatest or the least eigenvalue. When such is the case, iteration may be a convenient method of solution.

The general equation under consideration is (7.9), which is here reproduced as (7.53):

$$[A][X] = \lambda[B][X] \tag{7.53}$$

This could have been equally well written as

$$[B^{-1}][A][X] = \lambda[B^{-1}][B][X]$$

$$[B^{-1}A][X] = \lambda[X] \tag{7.54}$$

or

$$[A^{-1}][A][X] = \lambda[A^{-1}][B][X]$$

$$[X] = \lambda[A^{-1}B][X] \tag{7.55}$$

By noting that λ is a scalar quantity, Eq. (7.55) can also be written as

$$[A^{-1}B][X] = \frac{1}{\lambda}[X] \tag{7.55a}$$

In essence, the iterative procedure for determining eigenvalues is based on the condition that should a trial vector ${}_k[X]$ be assumed, an approximate eigenvalue and a second trial eigenvector ${}_{k+1}[X]$ can be determined from either Eq. (7.54) or Eq. (7.55*a*). It should be noted, however, that, as is true in all iteration methods, convergence is the real problem at issue. A strong diagonal set is a sufficient condition.

For illustration of the general method of solution under discussion, consider the vibration problem given as Example 7.3. In the matrix form of

(7.54), Eq. (7.35) can be written as

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.56}$$

The equation corresponding to (7.54) is

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.57}$$

In the terms of Eq. (7.55*a*), this same set can be written as

$$\begin{bmatrix} +0.60 & +0.20 \\ +0.20 & +0.40 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.58}$$

Equation (7.57) will first be considered, and as a starting point for the iteration procedure it will be assumed that the eigenvector is

$${}_0[X] = {}_0\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ +1.0 \end{bmatrix} \tag{7.59}$$

In general, convergence will be facilitated by choosing as an initial trial eigenvector the value +1.0 in the position corresponding to the strongest element in the $[A]$ matrix, while the others are kept zero.

Carrying out the matrix multiplication on the left-hand side,

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} 0 \\ +1.0 \end{bmatrix} = \begin{bmatrix} -1.0 \\ +3.0 \end{bmatrix} \tag{7.60}$$

If this column vector is now altered by removing the scalar quantity +3 (that is, by normalizing with respect to the a_2 element), it is observed that the form is exactly that of the right-hand side of Eq. (7.57). That is,

$$+3 \begin{bmatrix} -0.33 \\ +1.00 \end{bmatrix} = {}_1\lambda \, {}_1\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.61}$$

At this stage, then, a first approximation to one of the eigenvalues is ${}_1\lambda = +3$. The next trial value for the column vector of a's would be $a_1 = -0.33$ and $a_2 = +1.00$. A few of the trials are carried out below.

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} -0.33 \\ +1.00 \end{bmatrix} = \begin{bmatrix} -1.66 \\ +3.33 \end{bmatrix} = +3.33 \begin{bmatrix} -0.50 \\ +1.00 \end{bmatrix} = {}_2\lambda \, {}_2\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} -0.50 \\ +1.00 \end{bmatrix} = \begin{bmatrix} -2.00 \\ +3.50 \end{bmatrix} = +3.50 \begin{bmatrix} -0.572 \\ +1.000 \end{bmatrix} = {}_3\lambda \, {}_3\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} -0.572 \\ +1.000 \end{bmatrix} = \begin{bmatrix} -2.143 \\ +3.572 \end{bmatrix} = +3.572 \begin{bmatrix} -0.601 \\ +1.000 \end{bmatrix} = {}_4\lambda \, {}_4\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\begin{bmatrix} +2 & -1 \\ -1 & +3 \end{bmatrix} \begin{bmatrix} -0.601 \\ +1.000 \end{bmatrix} = \begin{bmatrix} -2.201 \\ +3.601 \end{bmatrix} = +3.601 \begin{bmatrix} -0.612 \\ +1.000 \end{bmatrix} = {}_5\lambda \, {}_5\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Table 7.1 summarizes the first nine iterations. It should be noted that the iteration process does converge to the correct larger eigenvalue with corresponding components for the eigenvector. [See Eqs. (7.52).] In general, *the iteration procedure using Eq. (7.54)*, if it converges, *will yield the largest eigenvalue.*

Table 7.1

Iterations:	1	2	3	4	5	6	7	8	9
λ	+3.00	+3.33	+3.50	+3.572	+3.601	+3.612	+3.614	+3.617	+3.618
a_1	−0.33	−0.50	−0.572	−0.601	−0.612	−0.614	−0.617	−0.618	−0.618
a_2	+1.00	+1.00	+1.000	+1.000	+1.000	+1.000	+1.000	+1.000	+1.000

Consider now the same problem, but expressed in the form of Eq. (7.58). As an initial trial assume $a_1 = +1.0$ and $a_2 = 0$.

$$\begin{bmatrix} +0.60 & +0.20 \\ +0.20 & +0.40 \end{bmatrix} \begin{bmatrix} +1.0 \\ 0 \end{bmatrix} = \begin{bmatrix} +0.60 \\ +0.20 \end{bmatrix} = +0.60 \begin{bmatrix} +1.00 \\ +0.33 \end{bmatrix} = \frac{1}{{}_1\lambda_1} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

The solution is summarized in Table 7.2. Here again the iteration converges. However, this time it is to the lower eigenvalue. In general, *the iteration procedure using Eq. (7.55a)*, if it converges, *will yield the smallest eigenvalue.*

Table 7.2

Iterations:	1	2	3	4	5	6	7	8
$1/\lambda$	+0.60	+0.667	+0.700	+0.714	+0.720	+0.722	+0.723	+0.723
λ	+1.667	+1.500	+1.429	+1.400	+1.390	+1.385	+1.382	+1.382
a_1	+1.00	+1.000	+1.000	+1.000	+1.000	+1.000	+1.000	+1.000
a_2	+0.33	+0.500	+0.572	+0.600	+0.612	+0.615	+0.617	+0.618

It should here be noted that iterative methods can be developed for determining intermediate eigenvalues. However, it is considered that these are beyond the scope of this introductory work.

7.4 *Modification of the Coefficient Matrix* $[B]$

Both in the exact expansion of the determinant and in the iterative method, a solution can be facilitated if $[B]$ can be reduced to the identity matrix or at least to that case where the leading diagonal elements are all 1.

When the $[B]$ *matrix is a diagonal matrix* of the form

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} b_{11} & 0 & \cdots & 0 \\ 0 & b_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ 0 & 0 & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{7.62}$$

a direct division of each equation by the b_{ii} term will give

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \lambda[I] \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{7.63}$$

where

$$\alpha_{ij} = \frac{a_{ij}}{b_{ii}} \tag{7.64}$$

The determinant to be solved is

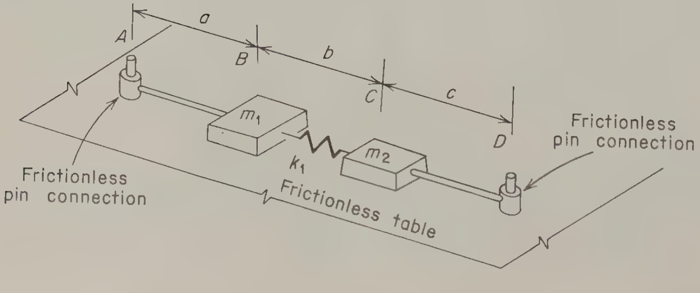
$$\begin{vmatrix} \alpha_{11} - \lambda & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} - \lambda & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} - \lambda \end{vmatrix} = 0 \tag{7.65}$$

This type of problem frequently occurs in lumped-parameter vibration problems.

Prob. 7.1: Determine the eigenvalues and eigenvectors for the following set of equations:

$$\begin{bmatrix} -12 & -24 & -37 \\ 0 & -24 & -32 \\ -48 & -24 & +36 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda[I] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Figure 7.6



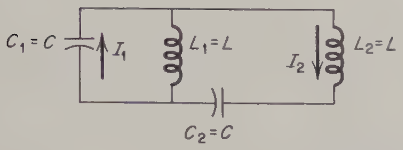


Figure 7.7

Prob. 7.2: By iteration, determine the largest and smallest eigenvalues for the following set of equations.

$$\begin{bmatrix} +4 & +2 & 0 \\ +2 & +10 & +4 \\ 0 & +4 & +5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda[I] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Prob. 7.3: Two masses m_1 and m_2, assumed to be resting on a frictionless table, are connected as shown in Fig. 7.6. The links connecting points A and B and points C and D are assumed to be inextensible. The spring between B and C is presumed to have a stiffness of k_1. Write in matrix form the equations governing the determination of the natural frequencies of vibration of the system.

Prob. 7.4: A column of uniform cross section, subjected to a vertical load at its uppermost point, has the following boundary conditions:

At $x = 0$: $\quad y = 0 \quad$ and $\quad y'' = 0$
At $x = L$: $\quad y = 0 \quad$ and $\quad y' = 0$

Using Eq. (7.43), determine the buckling loads.

Prob. 7.5: Assuming that the currents at resonance can be expressed as

$$I_1 = a_1 \sin(\omega t + \phi) \qquad I_2 = a_2 \sin(\omega t + \phi)$$

and the governing differential equations are of the form

$$L_i \frac{d^2 I_i}{dt^2} + \frac{I_j}{C_j} = 0 \qquad (i = 1, 2)$$

where I_i is the current in L_i, and I_j is the current in C_j, determine the resonant frequencies for the circuit given in Fig. 7.7.

Prob. 7.6: Two masses m_1 and m_2 are suspended from springs as shown in Fig. 7.8. The springs have stiffnesses of k_1 and k_2. What is the characteristic polynomial governing the eigenvalue solution to this problem?

Prob. 7.7: Determine the characteristic equations for the following sets of equations:

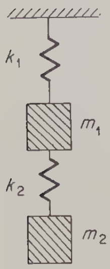


Figure 7.8

$$(a)\quad \begin{bmatrix} +2 & -1 & 0 & 0 \\ -1 & +2 & -2 & 0 \\ 0 & -2 & +4 & -4 \\ 0 & 0 & -4 & +6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \lambda \begin{bmatrix} +1 & 0 & 0 & 0 \\ 0 & +2 & 0 & 0 \\ 0 & 0 & +2 & 0 \\ 0 & 0 & 0 & +3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$(b)\quad \begin{bmatrix} +2 & -1 & 0 & 0 \\ -1 & +2 & -1 & 0 \\ 0 & -1 & +4 & -2 \\ 0 & 0 & -2 & +4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \lambda \begin{bmatrix} +1 & 0 & 0 & -1 \\ 0 & +2 & +1 & 0 \\ 0 & +1 & +2 & +1 \\ -1 & 0 & +1 & +3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

7.5 Computer Solutions

From the foregoing examples and problems, it should be evident that for larger-order systems, the numerical work associated with the direct algebraic solution of the normal eigenvalue problem [Eq. (7.11)] can become prohibitive. Iterative methods, on the other hand, can be numerically efficient for these types of problems provided only the lowest or highest eigenvalues are desired. For the determination of intermediate eigenvalues using iteration, it is necessary to develop methods for the removal from any assumed, or computed, eigenvector that part which corresponds to the higher (or lower) eigenvalues. These procedures can become quite complex. It should also be evident that many "real problems" contain the eigenvalue as a function of trigonometric or other functions. The algebraic methods that were discussed earlier in this chapter are not readily applied to these situations. For these and other reasons, many eigenvalue problems are evaluated by direct, systematic, trial-and-error procedures. It is this method that has been chosen for illustration.

The general problem type that will be considered is defined by the matrix equations (7.66) and (7.67).

$$[[A] - \lambda[B]][X] = 0 \tag{7.66}$$

$$[[A] - [B(\lambda)]][X] = 0 \tag{7.67}$$

In Eq. (7.67), the symbol $[B(\lambda)]$ has been used to indicate that the matrix $[B]$ can be any function of the unknown λ.

It was demonstrated earlier in this chapter that eigenvalue solutions correspond to those particular λ's for which

$$D = |A - \lambda B| = 0 \tag{7.68}$$

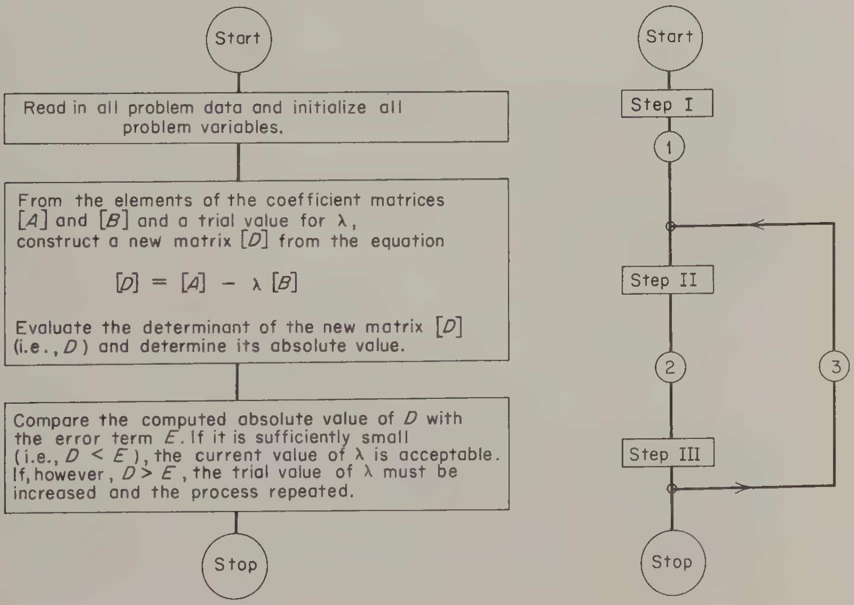
or

$$D = |A - B(\lambda)| = 0 \tag{7.69}$$

In this section, the general computational method that will be used to obtain computer solutions is as follows: A trial value of λ is first selected. With this, the determinant D, defined by either Eq. (7.68) or Eq. (7.69), is evaluated. Should $D = 0$, the trial value of λ corresponds to one of the eigenvalues. Should D be other than zero, a new λ is selected and D is again determined. If D is again not equal to zero, a comparison is made with the results of the previous trials and, based on these comparisons, a new λ is selected for trial. It should be noted that such a procedure can be made to give all the real eigenvalues contained within a given preselected range of λ's.

Seldom will a λ be found for which the value of the determinant is exactly zero. More often, a λ will be found for which D will be "sufficiently close to zero." It is necessary, then, to define at the outset the magnitude of the error to zero in the absolute value of the determinant that will be allowed. It will also be necessary, for automatic computation, to define the original increments in the trial values of the eigenvalues.

Figure 7.9



Read in all problem data and initialize all problem variables.

From the elements of the coefficient matrices $[A]$ and $[B]$ and a trial value for λ, construct a new matrix $[D]$ from the equation

$$[D] = [A] - \lambda [B]$$

Evaluate the determinant of the new matrix $[D]$ (i.e., D) and determine its absolute value.

Compare the computed absolute value of D with the error term E. If it is sufficiently small (i.e., $D < E$), the current value of λ is acceptable. If, however, $D > E$, the trial value of λ must be increased and the process repeated.

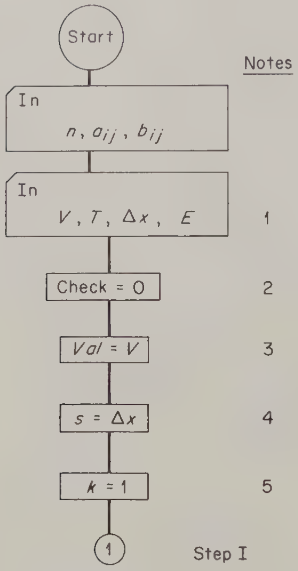


Figure 7.10

In broad terms, a flow diagram that will determine all the real eigenvalues contained within a preselected range of λ's is given as Fig. 7.9. The detailed diagrams for each of the parts are shown in Figs. 7.10 to 7.12. A complete flow diagram is contained in Fig. 7.13.

STEP I

Read in all problem data and initialize all problem variables.

A detailed flow diagram that will carry out these instructions is given in Fig. 7.10. The notes indicated in Fig. 7.10 have the following significance:

1. The variables shown in this box represent, in the order given, the following:

   V = any trial value for an eigenvalue (λ)
   T = maximum possible value of (λ) that is of interest
   Δx = original increment(s) in V
   E = error term which defines acceptable deviation from zero of absolute value of determinant

2. Since an incremental procedure is to be used, *Check* has been set aside to hold the last calculated value of the determinant. At each step, the new value of the determinant is compared with *Check*. When *Check* $< E$, an acceptable eigenvalue has been established.
3. *Val* is used to store the value of V.

4. s is a variable which defines the current increment to be taken. As a solution is approached, s will become successively smaller.
5. k is a fixed-point variable which defines the eigenvalue that has been found.

STEP II

From the elements of the coefficient matrices $[A]$ and $[B]$, and a trial value for λ, construct a new matrix $[D]$ from the equation

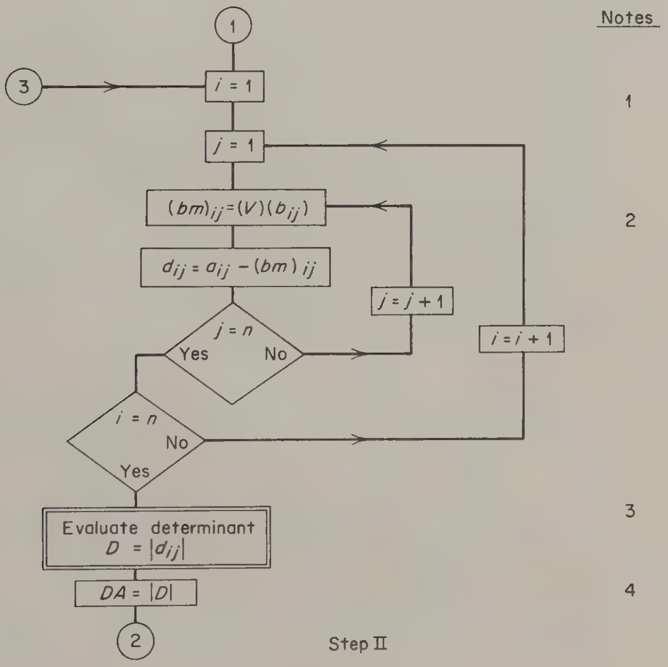
$$[D] = [A] - \lambda[B]$$

Evaluate the determinant of this new matrix $[D]$, that is, D, and determine its absolute value.

The following are noted in Fig. 7.11:

1. Point ③ is the position through which control returns to evaluate any trial determinant.
2. To eliminate the problem of destructive read-in, $(bm)_{ij}$ and d_{ij} are set up as new variables.
3. It is presumed that a *determinant-evaluation subroutine* is available (see Fig. 3.7).

Figure 7.11



4. *DA* is the absolute value of the trial determinant *D*.

STEP III

Compare the computed absolute value of *D* with the error term *E*. If it is sufficiently small, that is, $D < E$, the current value of λ is acceptable. If, however, $D > E$, the trial value of λ must be increased and the process repeated.

A flow diagram for carrying out Step III is shown in Fig. 7.12. The indicated notes correspond to the following:

1. If *DA* is found to be equal to or less than the allowable error *E*, *V* is an acceptable eigenvalue. Such solutions are put out as Eig_k. The variables are next reset, and control returns to step ③. If the absolute value of the determinant is greater than the allowed error, the value of the trial eigenvalue *V* must be increased and a new determinant evaluated.
2. If $V = Val$, that is, the first assumed trial eigenvalue is in fact an acceptable solution, the only course of action is to increase *V* by *s* and proceed to the determination of a second eigenvalue. If, however, *V* was not equal to *Val*, this was not the first cycle and the sign of *D* must be considered to ensure that the root is always approached in the positive direction.
3. At this point, a test is required to determine whether *D* is positive or negative. The sign of *D* when compared with the sign of *Check* (the last previous computed value of *D*) determines whether the root has been passed or not. If such is found to be the case, *V* is reduced to the last preceding value, *s* is reduced by ½, and *D* is replaced by the previous *D*. If the eigenvalue under consideration has not been passed, *Check* is replaced by *D*, a test is made for the maximum value of *V*, and *V* is increased by *s*. Control then returns to ③.
4. If $k = 1$ when $V \geq T$, no value of λ exists between the assumed starting value of *V* and the terminal value *T*. If $k \neq 1$, all the desired eigenvalues have been determined.

A complete flow diagram for the determination of eigenvalues in the manner just described is given in Fig. 7.13 (pages 156–157).

Prob. 7.8: The flow diagram just developed was based on the assumption that the problem was formulated according to Eq. (7.68). What changes in the program would be required to facilitate solution of problems of the Eq. (7.67) type?

Prob. 7.9: Construct a flow diagram for the determination of the smallest eigenvalue, and its corresponding eigenvector, using iteration.

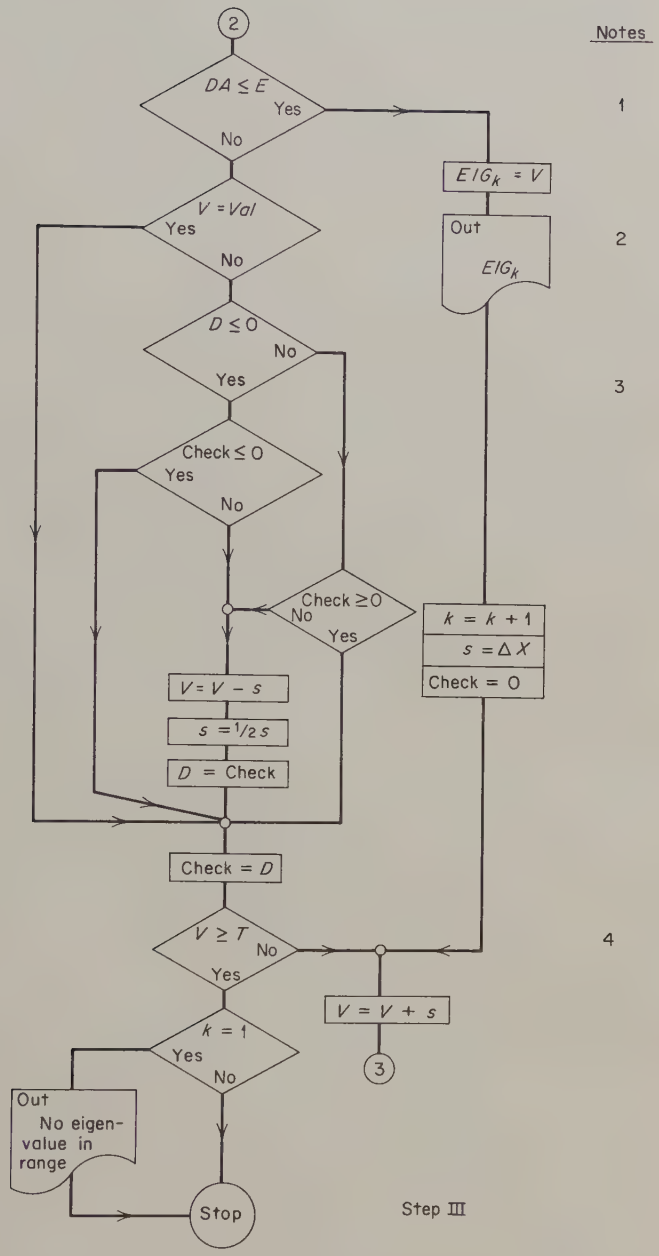

2
Notes
$DA \leq E$
Yes
No
1
$EIG_k = V$
$V = Val$
Yes
No
Out
EIG_k
2
$D \leq 0$
No
Yes
3
Check ≤ 0
Yes
No
Check ≥ 0
No
Yes
$k = k + 1$
$s = \Delta X$
Check = 0
$V = V - s$
$s = 1/2\, s$
D = Check
Check = D
$V \geq T$
No
Yes
4
$V = V + s$
3
$k = 1$
Yes
No
Out
No eigenvalue in range
Stop
Step III

Figure 7.12

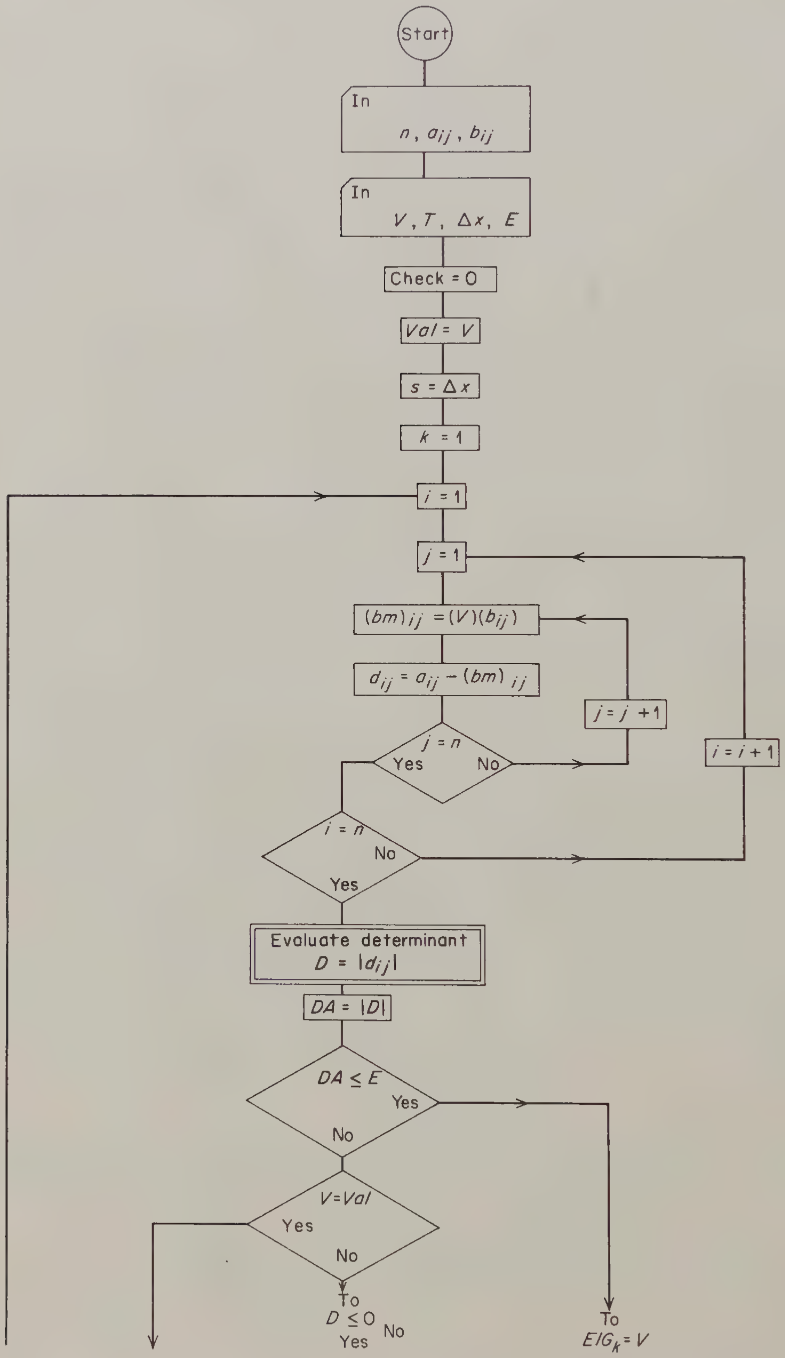


Figure 7.13 Flow diagram for determination of eigenvalues.

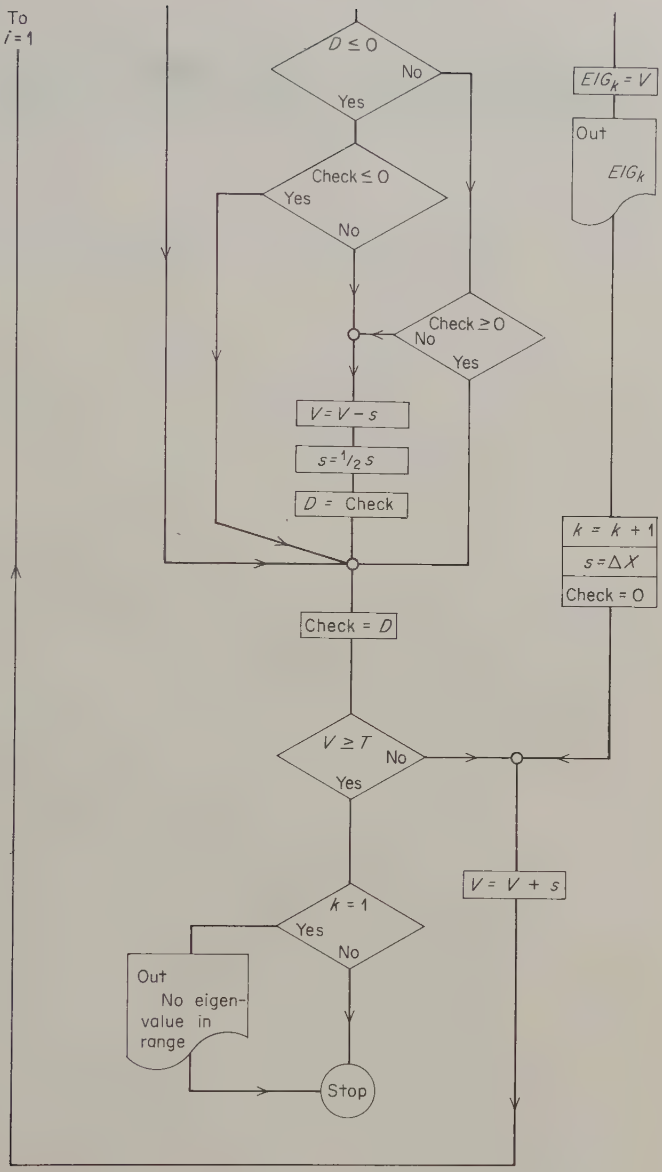


Figure 7.13 *(Continued)*

nonlinear equations

8

8.1 Introduction

By definition nonlinear equations are assumed to be those which contain powers or products of the variable(s) and/or transcendental functions. It is further presumed that transcendental encompasses all nonalgebraic situations, such as sin x, cos x, e^x, ln x, and $J_n(x)$.

Two problem types are of primary concern. The first is the evaluation of the "roots" of a given equation. The second is the determination of those particular values of the variables which simultaneously satisfy a given set of nonlinear equations.

Except for very special cases, no general algebraic methods are available for the solution of these types of problems. Under certain situations it may not even be possible to establish mathematically whether a solution (or solutions) does or does not exist. For these and other reasons, most solutions of nonlinear equations are obtained by iterative methods. As will be demonstrated later, to ensure convergence of these methods it is necessary that a reasonable approximation to the solution in question be known prior to iterating.

It should be understood at the outset that roots of nonlinear equations may be real or complex. While the major part of this chapter is concerned with methods for obtaining "real-root solutions," Sec. 8.5 discusses one method for obtaining complex roots of nonlinear algebraic equations.

Solution of problems of the type considered in this chapter normally consists of two phases: obtaining a first approximation, and improving the accuracy of the solution.

8.2 Methods for Obtaining First Approximations

8.2.1 Graphical Methods

One of the most frequently used scanning methods to establish first approximations to the solutions in question is that of sketching the curve(s) in the region desired. For example, consider the determination of the smallest real positive eigenvalue for a column fixed at its base and pin-connected at its uppermost point. The equation for which roots must be found is

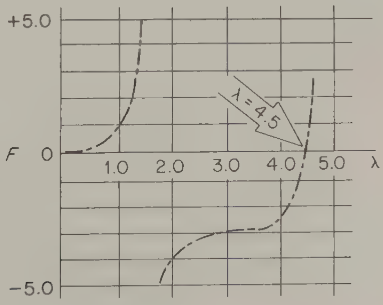
$$F = \tan \lambda - \lambda = 0 \tag{8.1}$$

Graphically, F can be plotted (or sketched) as a function of λ, as shown in Fig. 8.1. From this figure, it is evident that a first approximation to a positive-value solution other than zero would be $\lambda = +4.5$. It should be noted that while the general shape of F may (or may not) be readily apparent for this one particular given equation, such will not in general be true. Even for this case, possibly an easier scanning could have been realized by introducing a third variable z and separating the given equation (8.1) into the two equations

$$z = \tan \lambda \qquad z = \lambda \tag{8.2}$$

A simultaneous solution of these more elementary equations, the general shapes and magnitudes of which are known, corresponds to the actual solution of Eq. (8.1). The curves are shown in Fig. 8.2. Again, $\lambda = +4.5$.

Figure 8.1



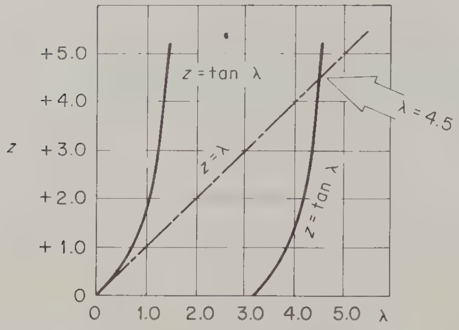


Figure 8.2

Approximations to the roots of most nonlinear equations can be obtained in this same manner. Consider the cubic equation

$$F = x^3 - 2x^2 + 2x - 1 = 0 \tag{8.3}$$

A direct plot of this equation is shown in Fig. 8.3. To facilitate solution, the original Eq. (8.3) can be rewritten in terms of a new variable y, where, by definition,

$$x = y + h \tag{8.4}$$

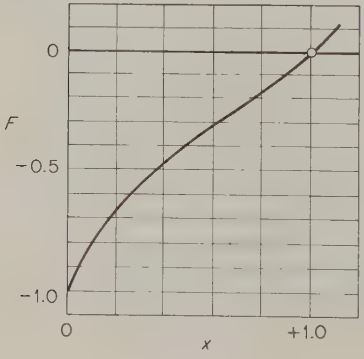
and h is a constant to be chosen at a later stage. Equation (8.3) can therefore be written as

$$(y + h)^3 - 2(y + h)^2 + 2(y + h) - 1 = 0 \tag{8.5}$$

or

$$y^3 + y^2(3h - 2) + y(3h^2 - 4h + 2) + (h^3 - 2h^2 + 2h - 1) = 0 \tag{8.6}$$

Figure 8.3



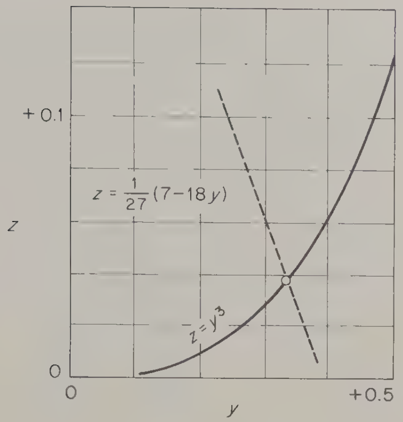


Figure 8.4

By selecting $h = +\frac{2}{3}$, it is noted that the coefficient of y^2 can be made to equal zero. The resulting equation is

$$y^3 + y(\tfrac{2}{3}) - \tfrac{7}{27} = 0 \tag{8.7}$$

which can be readily separated into the two more elementary equations

$$z = y^3 \qquad z = \tfrac{1}{27}(7 - 18y) \tag{8.8}$$

These two curves are sketched in Fig. 8.4. The solution is $y = +\frac{1}{3}$, or

$$x = y + h = \tfrac{1}{3} + \tfrac{2}{3} = +1.0 \tag{8.9}$$

Additional information is also evident from Fig. 8.4. Since for plus values of y, y^3 is a continuously increasing function; since for negative values, it has an increasing negative value; and since $z = \frac{1}{27}(7 - 18y)$ is a linear function of y, only one real root of Eq. (8.3) can be found.

As a third illustration of the graphical determination of first approximations to nonlinear equations, consider the two simultaneous nonlinear equations

$$\begin{aligned} f(x,y) &= 3y^3 - 2x^3 - 10x^2 + 1 = 0 \\ g(x,y) &= y^3 - 3x^2 + 5y - 5 = 0 \end{aligned} \tag{8.10}$$

One obvious method of solution is to plot "contour maps" of each equation and superimpose them. By noting that the first equation contains the variable y in only one term and that a similar situation for x exists in the second equation, Eqs. (8.10) can be written as

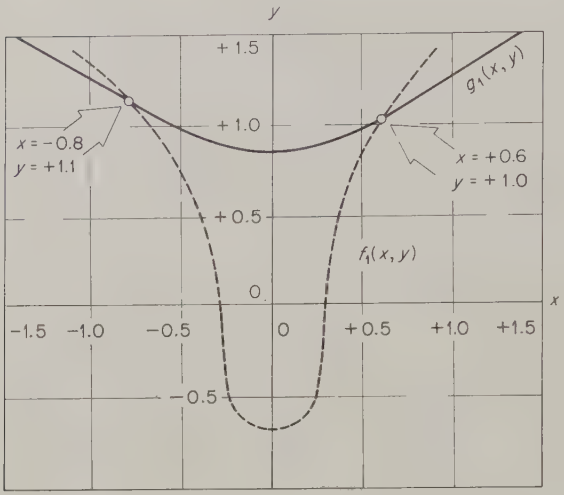


Figure 8.5

$$f_1(x,y) = y = \left[\frac{2x^2(x+5)-1}{3}\right]^{1/3}$$

$$g_1(x,y) = x = \left[\frac{y(y^2+5)-5}{3}\right]^{1/2} \tag{8.11}$$

These equations can be sketched as shown in Fig. 8.5. In the region shown, two solutions to the original set of equations exist.

8.2.2 Special Methods for Approximating the Real Roots of Nonlinear Algebraic Equations

The graphical procedure just described can be used to obtain approximate values of the roots of almost any nonlinear equation. However, when the equation is completely algebraic (i.e., a polynomial in x), there exist certain characteristics which can be of considerable aid in finding the approximate values of certain roots.

The algebraic equation in question is

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0 = 0 \tag{8.12}$$

In general, it has n roots. These may be real and distinct, real and repeating, or imaginary, in which case they would appear in couples of complex conjugate numbers.

Those real roots less than 1 in absolute value can be found by evaluating $f(x)$ at intervals of 0.2 or 0.1 in the range $-1 \le x \le +1$ and sketching the resulting curve. For real roots less than −1.0 or greater than +1.0,

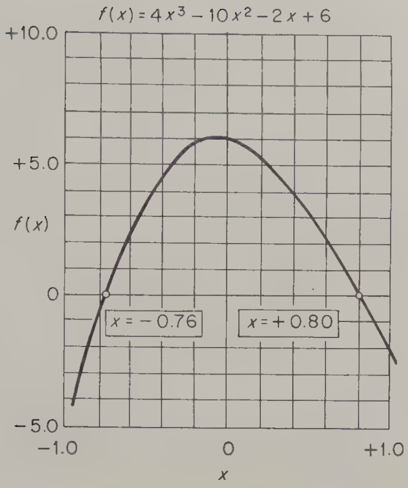


Figure 8.6

it is desirable to rewrite Eq. (8.12) in terms of a new variable z, where

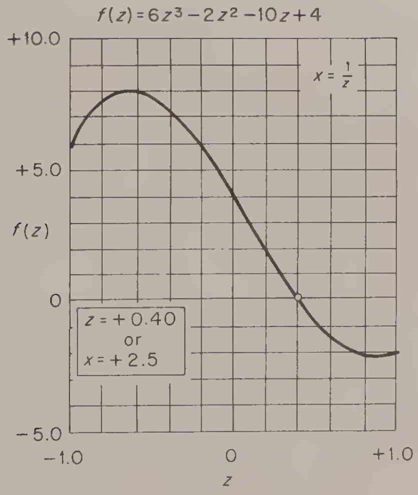
$$x = \frac{1}{z} \tag{8.13}$$

The equation is

$$f(z) = a_0z^n + a_1z^{n-1} + \cdots + a_{n-1}z + a_n = 0 \tag{8.14}$$

and the range of examination is $-1.0 \leq z \leq +1.0$.

Figure 8.7



For illustration of the general method, consider the polynomial

$$f(x) = 4x^3 - 10x^2 - 2x + 6 = 0 \tag{8.15}$$

Solutions are sketched in Figs. 8.6 and 8.7, and the approximate roots are

$$x = -0.76 \qquad x = +0.80 \qquad x = +2.50 \tag{8.16}$$

Alternative analytical methods for obtaining approximations to certain real roots are given by the following:

1. The *largest root* of a polynomial of the Eq. (8.12) type can be approximated by the larger of the roots of the following two equations:

$$a_n x + a_{n-1} = 0 \tag{8.17}$$

or

$$a_n x^2 + a_{n-1} x + a_{n-2} = 0 \tag{8.18}$$

For the previously considered polynomial, $4x^3 - 10x^2 - 2x + 6 = 0$, the equations corresponding to (8.17) and (8.18) are, respectively,

$$\begin{aligned} 4x - 10 &= 0 \\ 4x^2 - 10x - 2 &= 0 \end{aligned} \tag{8.19}$$

The root of the first of these is $x = +2.5$. The larger root of the second equation is $x = +2.69$.

2. *The smallest root* of a polynomial of the Eq. (8.12) type can be approximated by the roots of the following two equations:

$$\begin{aligned} a_1 x + a_0 &= 0 \\ \text{or} \qquad a_2 x^2 + a_1 x + a_0 &= 0 \end{aligned} \tag{8.20}$$

For Eq. (8.15), these are, respectively,

$$\begin{aligned} -2x + 6 &= 0 \\ -10x^2 - 2x + 6 &= 0 \end{aligned} \tag{8.21}$$

The roots are $x = +3.00$, from the first of these equations, and $x = -0.98$ and $+0.58$, from the second. An approximation to the smallest root is therefore $x = -0.98$.

8.2.3 Relationships That Exist between Roots of Algebraic Equations

As a guide (but not foolproof), the number of possible positive and negative real roots of equations of the type defined by Eq. (8.12) can be determined from Descartes's rule of signs:

1. The number of positive real roots is either equal to the number of variations in sign of the coefficients of the equation or less than that by a positive even integer.

2. The number of negative real roots is either equal to the number of repetitions in sign of the coefficients of the equation or less than that by a positive even integer.

Zero coefficients are ignored in this count. For illustration, again consider the polynomial

$$f(x) = 4x^3 - 10x^2 - 2x + 6 = 0 \tag{8.22}$$

There are two sign changes ($+4$ to -10 and -2 to $+6$). Therefore, there should be either two or no positive roots of Eq. (8.22). Since there is only one sign repetition (-10 to -2), there should be one negative real root. These were shown in Figs. 8.6 and 8.7.

Another general rule concerning the roots of equations of this type can be stated. If n is odd (that is, the equation is of odd degree), there will be at least one real root and it will have a sign opposite to the sign of a_0/a_n. Again consider Eq. (8.22). Since it is of odd degree (cubic), there must exist at least one real root whose sign is opposite to $+4/+6$.

It is possible to develop expressions relating the coefficients of the various elements in a polynomial to the roots of that equation. To illustrate these, again consider the general equation (8.12), here repeated as Eq. (8.23).

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0 = 0 \tag{8.23}$$

This can be written as

$$f(x) = x^n + \alpha_{n-1}x^{n-1} + \alpha_{n-2}x^{n-2} + \cdots + \alpha_1 x + \alpha_0 = 0 \tag{8.24}$$

where

$$\alpha_{n-i} = \frac{a_{n-i}}{a_n} \tag{8.25}$$

If the roots of this equation (that is, $x_1, x_2, \ldots, x_n$) are known, Eq. (8.23) can equally well be written as

$$f(x) = (x - x_1)(x - x_2)(x - x_3)(\cdots)(x - x_n) = 0 \tag{8.26}$$

Expanding Eq. (8.26) yields the equation

$$f(x) = x^n - \Big(\sum_{i=1}^{n} x_i\Big) x^{n-1} + \Big(\sum_{\substack{i=1\\j=1}}^{n} x_i x_j\Big) x^{n-2} - \Big(\sum_{\substack{i=1\\j=1\\k=1}}^{n} x_i x_j x_k\Big) x^{n-3} + \cdots$$

$$(-1)^n(x_1x_2x_3x_4 \cdots x_n) = 0 \qquad \text{for } i \neq j \neq k \neq \cdots \tag{8.27}$$

Since by definition Eqs. (8.24) and (8.27) represent the same polynomial, the coefficients of each must match. That is,

$$\begin{aligned} \alpha_{n-1} &= -\sum_{i=1}^{n} x_i \\ \alpha_{n-2} &= +\sum_{\substack{j=1\\ j=1}}^{n} x_i x_j \\ \alpha_{n-3} &= -\sum_{\substack{i=1\\ j=1\\ k=1}}^{n} x_i x_j x_k \\ &\cdots\cdots\cdots\cdots \\ \alpha_0 &= (-1)^n (x_1 x_2 x_3 x_4 \cdots x_n) \end{aligned} \tag{8.28}$$

where $i \neq j \neq k \neq \cdots$.

8.2.4 Synthetic Division

In determining the real roots of polynomials, or for that matter in finding the numerical value of any given nonlinear algebraic function at a preselected value of the independent variable, a process known as synthetic division may be useful. It should be here stated also that the process can be used to establish the numerical values of the derivatives of the given function at the preselected point. The relationship that exists is defined by the following:

$$\begin{aligned} \text{Remainder of first synthetic division} &= f(b) \\ \text{Remainder of second synthetic division} &= \frac{1}{1!} f'(b) \\ \text{Remainder of third synthetic division} &= \frac{1}{2!} f''(b) \\ \text{Remainder of fourth synthetic division} &= \frac{1}{3!} f'''(b) \\ &\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{8.29}$$

where b is the preselected point.

For illustration of the general procedure to be followed, consider the cubic equation given in (8.30). (The algebraic expressions for the derivatives are tabulated also for future reference.) If it is assumed that the value of x in question is 0.6, that is, $x = +0.6$, the numerical values of the function and its derivatives at this point are those given in Eqs. (8.31).

$$
\begin{aligned}
f(x) &= x^3 + 5x^2 + 0x - 2 \\
f'(x) &= 3x^2 + 10x \\
f''(x) &= 6x + 10
\end{aligned} \tag{8.30}
$$

$$
\begin{aligned}
f(0.6) &= +0.02 \\
f'(0.6) &= +7.08 \\
f''(0.6) &= +13.6
\end{aligned} \tag{8.31}
$$

Consider, now, a long-division evaluation of the function, presupposing $x = +0.6$, or $x - 0.6 = 0$.

$$
\begin{array}{r|l}
\multicolumn{1}{r}{} & +\qquad x^2 + 5.6x \ + 3.36 \\ \cline{2-2}
x - 0.6 & x^3 + \quad 5x^2 + \quad 0x - 2 \\
\multicolumn{1}{r}{} & \underline{x^3 - 0.6x^2} \\
\multicolumn{1}{r}{} & \quad + 5.6x^2 + \quad 0x \\
\multicolumn{1}{r}{} & \quad \underline{+ 5.6x^2 - 3.36x} \\
\multicolumn{1}{r}{} & \qquad\qquad + 3.36x - 2.00 \\
\multicolumn{1}{r}{} & \qquad\qquad \underline{+ 3.36x - 2.02} \\
\multicolumn{1}{r}{} & \qquad\qquad\qquad\qquad + 0.02 = \text{remainder} = f(0.6)
\end{array} \tag{8.32}
$$

Note that the remainder exactly equals the tabulated value for $f(0.6)$ given in (8.31), which was obtained by direct substitution. Dividing the quotient of Eq. (8.32), that is, the quadratic

$$x^2 + 5.6x + 3.36$$

by $x - 0.6$ will give as a remainder the value of the first derivative $f'(0.6)$:

$$
\begin{array}{r|l}
\multicolumn{1}{r}{} & \qquad\qquad x + 6.2 \\ \cline{2-2}
x - 0.6 & +x^2 + 5.6x + 3.36 \\
\multicolumn{1}{r}{} & \quad \underline{x^2 - 0.6x} \\
\multicolumn{1}{r}{} & \qquad + 6.2x + 3.36 \\
\multicolumn{1}{r}{} & \qquad \underline{+ 6.2x - 3.72} \\
\multicolumn{1}{r}{} & \qquad\qquad\qquad + 7.08 = \text{remainder} = f'(0.6)
\end{array} \tag{8.33}
$$

Again, dividing the "new" quotient by $x - 0.6$,

$$
\begin{array}{r|l}
\multicolumn{1}{r}{} & \quad + 1 \\ \cline{2-2}
x - 0.6 & x + 6.2 \\
\multicolumn{1}{r}{} & \underline{x - 0.6} \\
\multicolumn{1}{r}{} & \quad + 6.8 = \text{remainder}
\end{array} \tag{8.34}
$$

According to Eq. (8.29) the remainder of this third synthetic division equals $\frac{1}{2}f''(b)$. Therefore,

$$\begin{aligned} \tfrac{1}{2}f''(0.6) &= +6.8 \\ f''(0.6) &= +13.6 \end{aligned} \tag{8.35}$$

A more convenient form for synthetic division is shown in (8.36). In this formulation, the x^n terms are omitted and only the coefficients listed. (It should be noted that zero coefficients must be included when a given powered term is absent from the polynomial.) The original polynomial $x^3 + 5x^2 - 2$ is written as

$$1 \qquad +5 \qquad +0 \qquad -2$$

In "short form," the first synthetic division is

$$\begin{array}{lrrrl}
 & 1 & +5 & +0 & -2 \\
x_1 = 0.6 & & +0.6 & +3.36 & +2.02 \\
\hline
 & 1 & +5.6 & +3.36 & \multicolumn{1}{|l}{+0.02 = f(x) = f(0.6)} \\
\cline{5-5}
\end{array} \tag{8.36}$$

This same procedure can be continued to establish the values of each of the derivatives. The complete set is given as (8.37):

$$\begin{array}{lrrrl}
 & 1 & +5 & +0 & -2 \\
x_1 = 0.6 & & +0.6 & +3.36 & +2.02 \\
\hline
 & 1 & +5.6 & +3.36 & \multicolumn{1}{|l}{+0.02 = f(x_1)} \\
\cline{5-5}
x_1 = 0.6 & & +0.6 & +3.72 & \\
\cline{1-4}
 & 1 & +6.2 & \multicolumn{2}{|l}{+7.08 = f'(x_1)} \\
\cline{4-5}
x_1 = 0.6 & & +0.6 & & \\
\cline{1-3}
 & 1 & \multicolumn{3}{|l}{+6.8 = \dfrac{1}{2!} f''(x_1) \qquad \text{or} \qquad f''(x_1) = +13.6} \\
\cline{3-5}
\end{array} \tag{8.37}$$

8.3 *Methods for Improving the Accuracy of a Trial Solution*

First approximations to the real roots of a given nonlinear equation having been obtained, it is desirable to consider certain of the more common methods for improving these trial values.

8.3.1 False-position Method

Assume, for example, that the function values $f(x)$ for two relatively close trial values of x are known and that one of the function values is positive while the other is negative. Obviously, at least one root will be contained

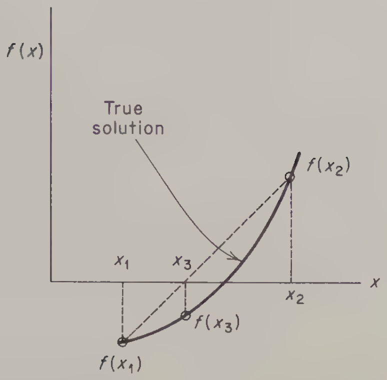


Figure 8.8

within this range of x. Figure 8.8 illustrates this situation. The two trial points correspond to x_1, $f(x_1)$ and x_2, $f(x_2)$.

An improved approximation to x corresponding to $f(x) = 0$ can be realized by linearly interpolating between these two points, as shown by the dotted lines in Fig. 8.8. The governing equation is

$$\frac{x - x_1}{f(x) - f(x_1)} = \frac{x_2 - x_1}{f(x_2) - f(x_1)} \tag{8.38}$$

where x and $f(x)$ correspond to any point on the line. For that particular point where $x = x_3$, and for which $f(x_3) = 0$,

$$x_3 = \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)} \tag{8.39}$$

For the case illustrated in Fig. 8.8, $f(x_3)$ also has a negative value. Therefore, the next linear interpolation would be between x_2 and x_3.

As a numerical example, consider the cubic equation (8.3), which is repeated here as Eq. (8.40):

$$f(x) = x^3 - 2x^2 + 2x - 1 = 0 \tag{8.40}$$

From Fig. 8.3, it is noted that $x = +1.0$; however, to illustrate this method of linear interpolation (or, as it is often referred to, the *method of false position*), it will be assumed that the solution is contained within the interval $+0.9 \leq x \leq +1.1$. By synthetic division or direct substitution, the following values are obtained:

$$\begin{aligned} x_1 &= +0.9 \qquad f(x_1) = f(0.9) = -0.091 \\ x_2 &= +1.1 \qquad f(x_2) = f(1.1) = +0.111 \end{aligned} \tag{8.41}$$

From Eq. (8.39) the value of x_3 for the "next trial" is

$$x_3 = \frac{(+0.9)(+0.111) - (+1.1)(-0.091)}{(+0.111) - (-0.091)} = \frac{+0.2000}{+0.202} = +0.990 \quad (8.42)$$

Substitution of this value in Eq. (8.40) yields

$$f(x_3) = f(0.99) = -0.0099 \quad (8.43)$$

Interpolation is therefore between x_2 and x_3, and x_4 is

$$x_4 = \frac{x_3 f(x_2) - x_2 f(x_3)}{f(x_2) - f(x_3)} = \frac{(+0.99)(+0.111) - (+1.1)(-0.0099)}{(+0.111) - (-0.0099)}$$

$$= +0.9992 \quad (8.44)$$

This process would be continued until the desired accuracy has been achieved.

8.3.2 Newton's Method of Tangents

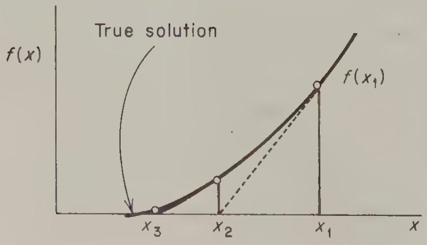
A second iterative method for improving the accuracy of a trial solution, known as *Newton's method of tangents* or just *Newton's method*, can also be used. In most cases it will result in faster convergence than will the false-position method. In addition, it requires as a starting situation a consideration of only one point.

Consider Fig. 8.9. It is assumed that x_1 is reasonably near the true solution and that $f(x_1)$ is the function value at that point. A second approximation can be realized by considering the tangent to the curve $f(x)$ at point x_1 and reducing the x_1 value accordingly. This is shown by the dotted line. The equation for the next approximation is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (8.45)$$

where $f'(x_n)$ is the first derivative of the function value evaluated at $x = x_n$. (The student is urged to verify the correctness of this equation

Figure 8.9



for positive or negative values of the function value as well as positive and negative first derivatives.)

Again, for the problem just considered,

$$f(x) = x^3 - 2x^2 + 2x - 1 = 0 \tag{8.46}$$

it will be assumed that $x_1 = +1.1$ is a reasonable first approximation. The first derivative with respect to x is

$$f'(x) = 3x^2 - 4x + 2 \tag{8.47}$$

The values to be used in Eq. (8.45) are therefore $f(1.1) = +0.111$ and $f'(1.1) = +1.23$.

$$x_2 = +1.1 - \frac{+0.111}{+1.23} = +1.1 - 0.0902 = +1.0098 \tag{8.48}$$

Several cycles of approximation are given in Table 8.1.

Table 8.1

Approx (n):	1	2	3
x_n	+1.1	+1.0098	+1.0001
$f(x_n)$	+0.111	+0.0099	
$f'(x_n)$	+1.23	+1.0198	
$\dfrac{f(x_n)}{f'(x_n)}$	+0.0902	+0.0097	

8.3.3 Newton's Second-order Method

A third method for achieving even faster convergence is that known as *Newton's second-order method.* If it is assumed that the function $f(x)$ varies continuously over the region in question, a Taylor series can be written with respect to a given point x_n. The function value at a new point h away from x_n (that is, at x_{n+1}) is defined by

$$f(x_{n+1}) = f(x_n) + f'(x_n)h + \frac{f''(x_n)}{2} h^2 + \cdots \tag{8.49}$$

where

$$x_{n+1} = x_n + h \tag{8.50}$$

By using only those terms of the series which are shown in Eq. (8.49), and assuming that the new function value under these conditions is the correct solution, that is, $f(x_{n+1}) = 0$, the following is obtained:

$$f(x_n) + h\left[f'(x_n) + \frac{f''(x_n)}{2} h\right] = 0 \tag{8.51}$$

For the h contained within the bracketed term, it will be assumed that the first-order Newton method, repeated here as Eq. (8.52), will hold.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{8.52}$$

That is,

$$h = x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} \tag{8.53}$$

Substitution of Eq. (8.53) into Eq. (8.51) yields

$$f(x_n) + h\left[f'(x_n) - \frac{f''(x_n)}{2}\frac{f(x_n)}{f'(x_n)}\right] = 0 \tag{8.54}$$

or

$$\frac{1}{h} = -\frac{f'(x_n)}{f(x_n)} + \frac{1}{2}\frac{f''(x_n)}{f'(x_n)} \tag{8.55}$$

For the cubic $x^3 - 2x^2 + 2x - 1 = 0$, the solution using this method in the vicinity of $x = +1.1$ is given in Table 8.2.

Table 8.2

Approx (n):	1	2	3
x_n	+1.1	+1.001	−1.000
$f(x_n)$	+0.111	+0.0010	
$f'(x_n)$	+1.23	+1.0020	
$f''(x_n)$	+2.60	+3.0060	
$\frac{1}{h}$	−10.115	−1003.0	
h	−0.099	−0.0010	

As was mentioned earlier in this section, it is frequently advantageous to define algebraically the derivatives of the equations in question and from these determine the function value and the value of the derivative at an assumed point rather than to use synthetic division. A symbolic notation for cumulative multiplication also can be used to computational advantage. Consider the following equation:

$$f(x) = x^6 - 18x^5 + 81x^4 - 100x^3 + 40x^2 - 20x + 4 = 0 \tag{8.56}$$

The first and second derivatives are given as (8.57) and (8.58):

$$f'(x) = 6x^5 - 90x^4 + 324x^3 - 300x^2 + 80x - 20 \tag{8.57}$$

$$f''(x) = 30x^4 - 360x^3 + 972x^2 - 600x + 80 \tag{8.58}$$

Introducing the symbol], which means that everything to the left of the symbol is to be multiplied by the quantity immediately following it, the three equations (8.56) to (8.58) can be written as

$$f(x) = x - 18]\,x + 81]\,x - 100]\,x + 40]\,x - 20]\,x + 4 \tag{8.59}$$

$$f'(x) = +6]\,x - 90]\,x + 324]\,x - 300]\,x + 80]\,x - 20 \tag{8.60}$$

$$f''(x) = +30]\,x - 360]\,x + 972]\,x - 600]\,x + 80 \tag{8.61}$$

It should be noted that such a formulation is efficient for both calculators and computer and can be carried out on most desk calculators without copying intermediate entries. Solution will be sought in the region $+4.0 \leq x \leq +5.0$. The function values at these points are

$$f(+4.0) = +564.0 \qquad f(+5.0) = -1{,}596.0$$

Solutions using each of the three methods are given in Tables 8.3 to 8.5.

Table 8.3 Solution by False Position

Approx (n):	1	2	3	4	5	6
x	+4.0	+5.0	+4.26	+4.38	+4.43	+4.435
$f(x_i)$	+564.0	−1,596.0	+313.2	+136.6	+50.1	

Table 8.4 Solution by First-order Newton's Method

Approx (n):	1	2	3	4	5
x_n	+4.0	+4.855	+4.551	+4.464	+4.457
$f(x_n)$	+564.0	−1,048.32	−193.41	−13.441	+0.011
$f'(x_n)$	−660.0	−3,444.85	−2,223.48	−1,922.38	
$\dfrac{f(x_n)}{f'(x_n)}$	−0.855	+0.304	+0.087	+0.007	

Table 8.5 Solution by Second-order Newton's Method

Approx (n):	1	2	3	4
x_n	+4.0	+4.152	+4.390	+4.459
$f(x_n)$	+564.0	+437.7	+120.0	−3.764
$f'(x_n)$	−3,540.0	−1,012.0	−1,642.8	
$f''(x_n)$	−2,128.0	−1,786.8	−2,416.7	
h	+0.152	+0.238	+0.069	

8.3.4 Direct Iteration

Direct iteration can also be used to improve the accuracy of a trial solution for the real roots of a nonlinear algebraic equation. In essence, this procedure requires that

$$f(x) = 0 \tag{8.62}$$

be replaced by

$$x = \phi(x) \tag{8.63}$$

Iteration is accomplished by substituting into the $\phi(x)$ function the values of a trial x; that is, the $(k + 1)$st trial value of x is obtained by substituting the kth trial value of x in $\phi(x)$.

$${}_{k+1}x = \phi({}_kx) \tag{8.64}$$

To facilitate convergence, it is necessary that the absolute value of the first derivative with respect to x of the function $\phi(x)$ be less than 1.0. Since in general there will be several possible $\phi(x)$'s that will satisfy this condition, the function whose derivative has an absolute value which is least in the range in question will be the most suitable.

Consider again, for illustration, the cubic equation (8.46), here repeated as Eq. (8.65):

$$x^3 - 2x^2 + 2x - 1 = 0 \tag{8.65}$$

Five elementary $\phi(x)$ functions can be found for this equation. These are given as Eqs. (8.66) through (8.70).

$$2x = -x^3 + 2x^2 + 1 \qquad \text{or} \qquad \phi_1(x) = \frac{-x^3 + 2x^2 + 1}{2} \tag{8.66}$$

$$2x^2 = x^3 + 2x - 1 \qquad \text{or} \qquad \phi_2(x) = \sqrt{\frac{x^3 + 2x - 1}{2}} \tag{8.67}$$

$$x(2)(x - 1) = x^3 - 1 \qquad \text{or} \qquad \phi_3(x) = \frac{x^3 - 1}{2(x - 1)} \tag{8.68}$$

$$x^3 = 2x^2 - 2x + 1 \qquad \text{or} \qquad \phi_4(x) = \sqrt[3]{2x^2 - 2x + 1} \tag{8.69}$$

$$x(x^2 - 2x + 2) = 1 \qquad \text{or} \qquad \phi_5(x) = \frac{1}{x^2 - 2x + 2} \tag{8.70}$$

Expressions for the derivatives of these functions with respect to x are given as Eqs. (8.71) through (8.75). Values have been calculated for $x = +1.1$.

$$\phi_1'(x) = \frac{\partial \phi_1(x)}{\partial x} = \frac{1}{2}(-3x^2 + 4x) \qquad \phi_1'(1.1) = +0.385 \tag{8.71}$$

$$\phi_2'(x) = \frac{\partial \phi_2(x)}{\partial x} = \frac{1}{4}\frac{3x_2 + 2}{\sqrt{x^3 + 2x - 1}} \qquad \phi_2'(1.1) = +0.885 \tag{8.72}$$

$$\phi_3'(x) = \frac{\partial \phi_3(x)}{\partial x} = \frac{1}{2}\frac{2x^3 + 3x^2 + 1}{(x-1)^2} \qquad \phi_3'(1.1) = +1.600 \tag{8.73}$$

$$\phi_4'(x) = \frac{\partial \phi_4(x)}{\partial x} = \frac{1}{3}\frac{4x - 2}{\sqrt[3]{(2x^2 - 2x + 1)^2}} \qquad \phi_4'(1.1) = +0.642 \tag{8.74}$$

$$\phi_5'(x) = \frac{\partial \phi_5(x)}{\partial x} = -\frac{2(x-1)}{(x^2 - 2x + 2)^2} \qquad \phi_5'(1.1) = -0.186 \tag{8.75}$$

From the above, $\phi_5(x)$ is the most desirable iterating function. Several cycles using it result in the following:

$$+1.1,\ +0.99,\ +1.000,\ +1.000,\ \ldots \tag{8.76}$$

The first three cycles using $\phi_3(x)$ are shown in (8.77). It should be remembered that this latter function does not satisfy the condition that $\phi'(x) < 1$, and therefore it is not a suitable iterating function.

$$+1.1,\ +1.655,\ +3.456,\ \ldots \tag{8.77}$$

8.4 *Real Roots of Transcendental Equations*

Transcendental equations, like algebraic equations, may exist as single or simultaneous equations. They may also have real or imaginary roots.

In general, all the methods described in the preceding parts of this chapter are directly applicable to the solution of transcendental equations. In addition, the transcendental elements of a given equation(s) can be represented, to the accuracy desired, in series form and thereby reduce the system to nonlinear algebraic equation(s). Table 8.6 summarizes a few of the more frequently referred to series.

Table 8.6

Function	Series
e^x	$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$
$\sin x$	$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$
$\cos x$	$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$
$\tan x$	$x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \cdots$

For illustration of the method of solution of transcendental equations by expansion in series form, consider the determination of the smallest positive root of Eq. (8.78).

$$f(x) = \frac{1-x}{x} - \tan x = 0 \tag{8.78}$$

By replacing $\tan x$ by its series (from Table 8.6), Eq. (8.78) can be written as

$$f(x) = \frac{1-x}{x} - \left(x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \cdots\right) = 0 \tag{8.79}$$

For an approximate value of the root, consider only the first term of the series, that is,

$$\frac{1-x}{x} - x = 0 \tag{8.80}$$

or

$$x^2 + x - 1 = 0 \tag{8.81}$$

which has roots

$$x = +0.56 \text{ and } -1.62 \tag{8.82}$$

A more exact answer can be obtained by considering the first three terms of the tangent series.

$$f(x) = \frac{1-x}{x} - \left(x + \frac{x^3}{3} + \frac{2x^5}{15}\right) = 0 \tag{8.83}$$

or

$$2x^6 + 5x^4 + 15x^2 + 15x = 15 \tag{8.84}$$

This can be written in the iterative form of Eq. (8.63) by factoring from the left-hand side of the equation an x, as shown in Eq. (8.85):

$$x = \phi(x) = \frac{15}{2x^5 + 5x^3 + 15x + 15} \tag{8.85}$$

Since the value of the derivative with respect to x of $\phi(x)$ at $x = +0.56$ is much less than 1.0, iteration should converge. Several cycles are as follows:

$$+0.56, +0.615, +0.587, +0.601, \ldots \tag{8.86}$$

Direct application of Newton's method to the equation in its original form can also be carried out to obtain a solution to this problem. The appropriate expressions are

$$\begin{aligned} f(x) &= \frac{1-x}{x} - \tan x = 0 \\ f'(x) &= -\frac{1}{x^2} - \sec^2 x \end{aligned} \tag{8.87}$$

By noting that $\sec^2 x = 1 + \tan^2 x$, the iterative form of the equation is

$$ {}_{k+1}x = {}_kx - \frac{{}_kx - {}_kx^2(1 + \tan {}_kx)}{-1 - {}_kx^2(1 + \tan^2 {}_kx)} \tag{8.88} $$

Table 8.7 summarizes several cycles of iteration. Graphically, the equations are plotted in Fig. 8.10. Only the region $-1.5 < x < +1.5$ is shown.

Table 8.7

k	0	1	2
x	$+0.56$	$+0.595$	$+0.596$
$\frac{f(x)}{f'(x)}$	-0.035	-0.001	

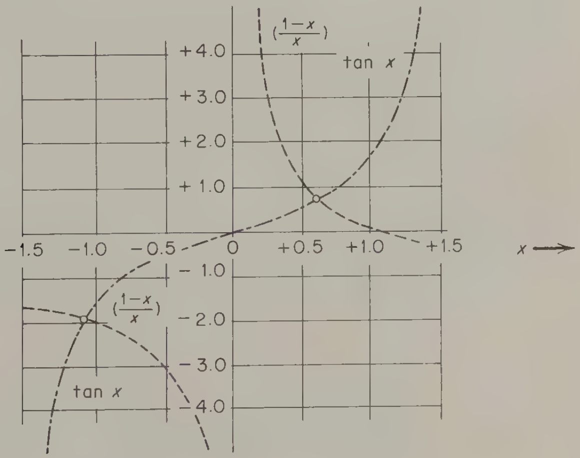
8.5 *Complex Roots of Nonlinear Algebraic Equations*

The general polynomial equation under consideration is that listed originally as Eq. (8.12) and here repeated as Eq. (8.89).

$$ f(x) = a_nx^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0 = 0 \tag{8.89} $$

It is assumed that both real and complex roots may exist.

Figure 8.10



By presupposing that the real roots have been established (for example, x_1, x_2, etc.), the original equation can be reduced in order by dividing synthetically from Eq. (8.89) the values $x - x_1$, $x - x_2$, ... The polynomial remaining after these expressions have been removed is

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0 = 0 \tag{8.90}$$

Complex roots must exist in conjugate pairs of the form

$$\zeta \pm i\xi \tag{8.91}$$

Therefore, the product

$$[x - (\zeta + i\xi)][x - (\zeta - i\xi)]$$

must be a factor of the equation in question. Expanding this relationship,

$$x^2 - 2\zeta x + (\zeta^2 + \xi^2) = 0 \tag{8.92}$$

Equation (8.92) is a quadratic in x having real coefficients.

Should it be possible to find a quadratic equation of the general form

$$x^2 + ux + v = 0 \tag{8.93}$$

which is an exact divisor of the original equation (8.89) [or (8.90)], the roots of that quadratic will represent two of the complex roots in question. The problem, then, is to develop a method for determining suitable values of u and v.

If it is assumed that certain trial values for u and v are known, the given polynomial $f(x)$ of Eq. (8.89) can be written as in Eq. (8.94).

$$f(x) = (x^2 + ux + v)(c_n x^{n-2} + c_{n-1}x^{n-3} + \cdots + c_3 x + c_2) + c_1 x + c_0 \tag{8.94}$$

For the correct values of u and v, *c_1 and c_0 will equal zero.*

The coefficients $c_n, c_{n-1}, \ldots, c_3, c_2$ in the reduced equation can be directly related to the original coefficients $a_n, a_{n-1}, \ldots, a_1, a_0$ and the trial values u and v. Expanding Eq. (8.94),

$$\begin{aligned} c_n x^n &+ (c_n u + c_{n-1})x^{n-1} + (c_n v + c_{n-1}u + c_{n-2})x^{n-2} \\ &+ (c_{n-1}v + c_{n-2}u + c_{n-3})x^{n-3} + \cdots + (c_3 v + c_2 u + c_1)x \\ &+ (c_2 v + c_0) = 0 \end{aligned} \tag{8.95}$$

A comparison of the coefficients is given in Table 8.8.

To illustrate the method of solution using the relationships of Table 8.8, consider the solution of the cubic equation

$$x^3 + x + 1 = 0 \tag{8.96}$$

For this case $a_3 = +1$, $a_2 = 0$, $a_1 = +1$, and $a_0 = +1$. *The object is to select u and v so that c_1 and c_0 approach zero.* While there can be developed iterative methods for the solution of the two-independent-variable type

Table 8.8

$a_n = c_n$	$c_n = a_n$
$a_{n-1} = c_n u + c_{n-1}$	$c_{n-1} = a_{n-1} - c_n u$
$a_{n-2} = c_n v + c_{n-1} u + c_{n-2}$	$c_{n-2} = a_{n-2} - c_{n-1} u - c_n v$
$a_{n-3} = c_{n-1} v + c_{n-2} u + c_{n-3}$	$c_{n-3} = a_{n-3} - c_{n-2} u - c_{n-1} v$
. . . .	. . . .
$a_1 = c_3 v + c_2 u + c_1$	$c_1 = a_1 - c_2 u - c_3 v$
$a_0 = c_2 v + c_0$	$c_0 = a_0 - c_2 v$

of problem, a trial-and-error approach will be demonstrated for this example. To facilitate investigation, the form of Tables 8.9 and 8.10 will be used.

Table 8.9

		Computed values of c_n					
		1	2	3	4	5	6
	u	−0.5	−0.8	−1.0	−0.5	−0.8	−1.0
	v	+1.5	+1.5	+1.5	+1.0	+1.0	+1.0
$a_3 = +1$	$c_3 = a_3$	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00
$a_2 = 0$	$c_2 = a_2 - c_3 u$	+0.50	+0.80	+1.00	+0.50	+0.80	+1.00
$a_1 = +1$	$c_1 = a_1 - c_2 u - c_3 v$	+0.25	+0.14	+0.50	+0.25	+0.64	+1.00
$a_0 = +1$	$c_0 = a_0 - c_2 v$	+0.25	+0.20	−0.50	+0.50	+0.20	0

The results of trials 1 through 9 are plotted in Fig. 8.11, where the ordinate is the function value of either c_1 or c_0, and the abscissa is u. Based on Fig. 8.11, trial 10 was carried out. By assuming that $c_1 = -0.01$ and $c_0 = -0.05$ are sufficiently close to zero, the complex roots of Eq. (8.96) will correspond to the roots of Eq. (8.97).

$$x^2 - 0.7x + 1.5 = 0 \tag{8.97}$$

That is,

$$x = +0.35 \pm \sqrt{(0.35)^2 - 1.5} = +0.35 \pm 1.17i \tag{8.98}$$

These same values could have been obtained more directly by noting the relationship between u and v and ζ and ξ [see Eqs. (8.92) and (8.93)].

Table 8.10

	Computed values of c_n			
	7	8	9	10
u	-0.5	-0.8	-1.0	-0.7
v	$+2.0$	$+2.0$	$+2.0$	$+1.5$
c_3	$+1.00$	$+1.00$	$+1.00$	$+1.00$
c_2	$+0.50$	$+0.80$	$+1.00$	$+0.70$
c_1	-0.50	-0.36	0	-0.01
c_0	0	-0.60	-1.00	-0.05

Since

$$-2\zeta = u = -0.70 \qquad \zeta = +0.35 \tag{8.99}$$

and

$$\zeta^2 + \xi^2 = v = +1.5 \qquad \text{or} \qquad \xi = +1.17 \tag{8.100}$$

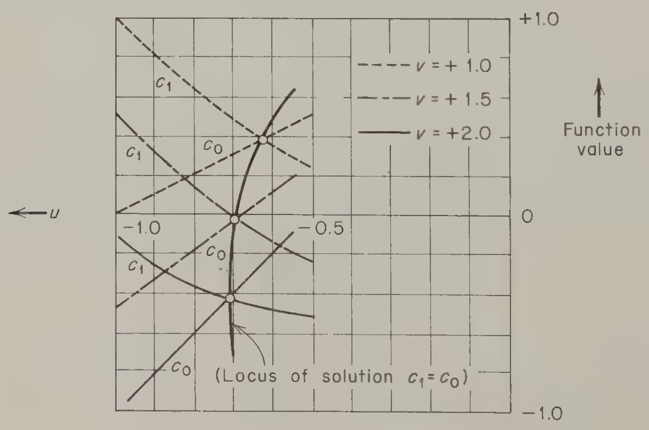
Therefore, from (8.91),

$$x = \zeta \pm i\xi = +0.35 \pm 1.17i \tag{8.101}$$

8.6 Real Roots of Simultaneous Nonlinear Algebraic Equations

The methods discussed earlier in this chapter for finding the roots of a single nonlinear algebraic equation in one independent variable are not directly adaptable to those cases where more than one equation is speci-

Figure 8.11



fied. The methods can, however, be modified to take care of these cases, as will be demonstrated in this section.

For illustration, the two-independent-variable set of two nonlinear equations defined by (8.102) will be considered.

$$f(x,y) = 0 \qquad g(x,y) = 0 \tag{8.102}$$

Larger-order systems would be examined in a similar manner. Two methods of solution will be considered: (1) Newton's method and (2) direct iteration.

8.6.1 Newton's Method

By assuming that a sufficiently close approximate solution $x = a_1$ and $y = b_1$ is known, Eqs. (8.102) can be expanded into Taylor series about these points. By using h and k to represent the changes in x and y, respectively, which need to be made in successive cycles of approximation, the expansion of Eqs. (8.102) can be written to a first-order approximation as

$$\begin{aligned} f(a_1 + h,\, b_1 + k) &= f(a_1,b_1) + hf_x(a_1,b_1) + kf_y(a_1,b_1) = 0 \\ g(a_1 + h,\, b_1 + k) &= g(a_1,b_1) + hg_x(a_1,b_1) + kg_y(a_1,b_1) = 0 \end{aligned} \tag{8.103}$$

where

$$\begin{aligned} f_x(a_1,b_1) &= \frac{\partial f(a_1,b_1)}{\partial x} \qquad & f_y(a_1,b_1) &= \frac{\partial f(a_1,b_1)}{\partial y} \\ g_x(a_1,b_1) &= \frac{\partial g(a_1,b_1)}{\partial x} \qquad & g_y(a_1,b_1) &= \frac{\partial g(a_1,b_1)}{\partial y} \end{aligned} \tag{8.104}$$

and

$$x = a_1 + h \qquad y = b_1 + k \tag{8.105}$$

Equations (8.103) can be solved simultaneously to define h and k.

$$h = \frac{gf_y - fg_y}{f_xg_y - g_xf_y} \tag{8.106}$$

$$k = \frac{fg_x - gf_x}{f_xg_y - g_xf_y} \tag{8.107}$$

where the function values and their derivatives are evaluated at $x = a_1$ and $y = b_1$.

For illustration of the procedure, consider the following two simultaneous nonlinear algebraic equations:

$$\begin{aligned} f(x,y) &= 3y^3 - 2x^3 - 10x^2 + 1 = 0 \\ g(x,y) &= y^3 - 3x^2 + 5y - 5 = 0 \end{aligned} \tag{8.108}$$

The various required derivatives are

$$f_x = -6x^2 - 20x \qquad g_x = -6x$$
$$f_y = +9y^2 \qquad g_y = +3y^2 + 5 \tag{8.109}$$

Table 8.11 gives the solution, assuming as approximations $x_1 = +0.6$ and $y_1 = +1.0$.

Table 8.11

Cycles:	1	2	3
x	+0.6	+0.606	+0.606
y	+1.0	+1.013	+1.013
f	−0.032	+0.001	
f_x	−14.160	−14.323	
f_y	+9.000	+9.235	
g	−0.080	+0.003	
g_x	−3.600	−3.636	
g_y	+8.000	+8.078	
h	+0.006	−0.00023	
k	+0.013	−0.00047	

8.6.2 Direct Iteration

For direct iteration instead of Newton's method, it will be assumed that the two equations can be written in the form

$$f(x,y) = x - \phi(x,y) = 0$$
$$g(x,y) = y - \psi(x,y) = 0 \tag{8.110}$$

By presupposing sufficiently close approximate solutions $x = a_1$ and $y = b_1$, iteration can proceed from the equalities

$${}_{k+1}x = \phi({}_kx,{}_ky)$$
$${}_{k+1}y = \psi({}_kx,{}_ky) \tag{8.111}$$

It should be noted that k is used here to designate the cycle of iteration, as was done in Eq. (8.64); it is not used to designate the change in y, as was the case in Eqs. (8.105).

Many functions $\phi(x,y)$ and $\psi(x,y)$ can be found for a given set of nonlinear equations. Guest† has shown that for convergence, it is desirable to select functions which satisfy the following conditions:

---

† J. Guest, On the Choice of Suitable Functions for Solving Non-linear Equations by Iterative Procedures, *Aeron. Res. Lab., Struct. Mater. Note* 262, Australian Defense Scientific Service, Melbourne, November, 1959.

$$\left|\frac{\partial \phi}{\partial x}\right|, \left|\frac{\partial \phi}{\partial y}\right|, \left|\frac{\partial \psi}{\partial x}\right|, \left|\frac{\partial \psi}{\partial y}\right| < 2$$

and

$$\left|\frac{\partial \phi}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial \phi}{\partial y}\frac{\partial \psi}{\partial x} - \frac{\partial \phi}{\partial x} - \frac{\partial \psi}{\partial y}\right| < 1 \qquad (8.112)$$

In addition, since the first of Eqs. (8.110) will essentially be used to define x while the second equation will determine y, it is desirable to have $|\partial\phi/\partial x|$ and $|\partial\psi/\partial y|$ as small as possible.

Consider the two simultaneous nonlinear algebraic equations (8.108) here repeated as Eqs. (8.113).

$$\begin{aligned} f(x,y) &= 3y^3 - 2x^3 - 10x^2 + 1 = 0 \\ g(x,y) &= y^3 - 3x^2 + 5y^2 - 5 = 0 \end{aligned} \qquad (8.113)$$

From these, by considering Eqs. (8.110), the following elementary functions can be written:

$$\begin{aligned} \phi_1 &= \sqrt{\frac{3y^3 - 2x^3 + 1}{10}} & \psi_1 &= \sqrt[3]{\frac{2x^3 + 10x^2 - 1}{3}} \\ \phi_2 &= \sqrt[3]{\frac{3y^3 - 10x^2 + 1}{2}} & \psi_2 &= \sqrt{3x^2 - 5y + 5} \\ \phi_3 &= \frac{3y^3 + 1}{2x^2 + 10x} & \psi_3 &= \frac{3x^2 - y^3 + 5}{5} \\ \phi_4 &= \sqrt{\frac{3y^3 + 1}{2x + 10}} & \psi_4 &= \frac{3x^2 + 5}{y^2 + 5} \\ \phi_5 &= \sqrt{\frac{y^3 + 5y - 5}{3}} & \psi_5 &= \sqrt{\frac{3x^2 + 5}{y} - 5} \end{aligned} \qquad (8.114)$$

Since ϕ_5 is a function only of y and ψ_1 varies only with x, one may be tempted to assume without checking that iteration using these functions will converge. Such is not the case.

$$\frac{\partial \phi_5}{\partial x} = 0 \qquad \frac{\partial \phi_5}{\partial y} = +2.31$$

$$\frac{\partial \psi}{\partial x} = +1.56 \qquad \frac{\partial \psi}{\partial y} = 0 \qquad (8.115)$$

$$\frac{\partial \phi}{\partial x}\frac{\partial \psi}{\partial y} - \frac{\partial \phi}{\partial y}\frac{\partial \psi}{\partial x} - \frac{\partial \phi}{\partial x} - \frac{\partial \psi}{\partial y} = -3.60$$

Table 8.12 summarizes the values of the various derivatives at $x = +0.6$ and $y = +1.0$.

Table 8.12

	$\frac{\partial\phi}{\partial x}$	$\frac{\partial\phi}{\partial y}$	$\frac{\partial\psi}{\partial x}$	$\frac{\partial\psi}{\partial y}$	R†
ϕ_1	−0.18	+0.75			
ϕ_2	−17.5	+13.1			
ϕ_3	−1.10	+1.34			
ϕ_4	−0.05	+0.67			
ϕ_5	0	+2.31			
ψ_1			+1.56	0	
ψ_2			+1.14	−1.58	
ψ_3			+0.72	−0.60	
ψ_4			+0.60	−0.34	
ψ_5			+1.04	−2.92	
ϕ_4, ψ_4					+0.01

† $R = \frac{\partial\phi}{\partial x}\frac{\partial\psi}{\partial y} - \frac{\partial\phi}{\partial y}\frac{\partial\psi}{\partial x} - \frac{\partial\phi}{\partial x} - \frac{\partial\psi}{\partial y}$

Functions ϕ_4 and ψ_4 are selected as the iterating functions. That is,

$$x = \sqrt{\frac{3y^3 + 1}{2x + 10}} \qquad y = \frac{3x^2 + 5}{y^2 + 5} \tag{8.116}$$

Several cycles of iteration are given in Table 8.13.

Table 8.13

Cycles:	0	1	2	3	4
x	+0.6	+0.598	+0.607	+0.603	+0.607
y	+1.0	+1.013	+1.008	+1.015	+1.010

Prob. 8.1: Using the false-position, Newton's first-order, and Newton's second-order methods, determine the smallest positive roots of the following equations. Carry the answers to three decimal places.

(a) $x^3 + 12x + 3 = 0$
(b) $x^3 + 3x + 2 = 0$
(c) $x^4 + x^3 + 6x^2 - x + 3 = 0$
(d) $x^3 + x^2 - 4x = 0$
(e) $x^3 + 3x - 3 = 0$
(f) $2x^5 + x^4 + 11x^3 - 8x^2 + 7x - 3 = 0$

Prob. 8.2: Determine all the roots of the following equations and check the results by using the relationships given in Eqs. (8.28).

(a) $x^3 + 3x - 3 = 0$
(b) $x^4 - 10x^3 + 35x^2 - 50x + 24 = 0$
(c) $x^4 - 4x^3 - 7x^2 + 22x + 24 = 0$

Prob. 8.3: Determine the smallest roots (in absolute value) of the following transcendental equations:

(a) $1 - x + x \sin x = 0$
(b) $6x_2 - e^x = 0$
(c) $3 \tan^2 \theta + 2 \sin \theta - \theta = 0$

Prob. 8.4: Determine the roots of the following sets of nonlinear algebraic equations in the range $0 \leq x \leq +1.0$ and $0 \leq y \leq +1.0$:

(a) $x^2 - 2xy + y^2 - 2x - 2y + 1 = 0$
$y^2 = x^3$
(b) $4yx^2 - 2x + y = 0$
$yx^2 + y - 1 = 0$
(c) $y^2(1 + x) = x^2(1 - x)$
$6x^3 - 2x^2 - 10x + 4 = 0$

Prob. 8.5: Find the conjugate complex roots of the following cubic equations:

(a) $x^3 + 4x + 2 = 0$
(b) $x^3 - 2x + 12 = 0$
(c) $x^3 - 6x^2 + 11x - 6 = 0$

Prob. 8.6: The determination of the buckling load of a column fixed at its base and pin-connected at its uppermost point requires the solution of the equation

$$\lambda = \tan \lambda$$

Determine the lowest three positive roots of this equation.

Prob. 8.7: Using both Newton's method and the method of direct iteration, determine to three decimal places the square roots of the following numbers:

(a) 2.000 (b) 23.676 (c) 61.384 (d) 147.361

Prob. 8.8: For the numbers given in Prob. 8.7, determine the cube roots by using Newton's second-order method. Carry the solutions to three decimal places.

Prob. 8.9: Given the equation

$$f(x) = \sqrt[3]{3 - x} = 0$$

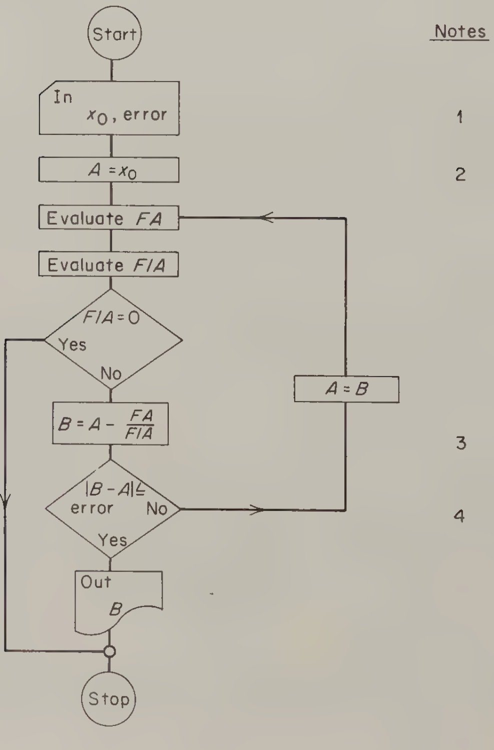
and assuming, as a starting value, that the root in question is in the vicinity of $x = +3.1$, carry out the first four iterations using Newton's first-order method. Discuss the convergence of the solution and indicate how this problem is different from those illustrated in the text.

8.7 Computer Solutions

Since, computationally, all the methods for determining the roots of nonlinear algebraic equations discussed in this chapter are of the same general type, only Newton's first-order method will be illustrated. The others would be handled in a similar manner.

It will be recalled from Sec. 8.3 that Newton's first-order method for

Figure 8.12



determining the roots of a given nonlinear algebraic equation requires the repeated application of Eq. (8.117).

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{8.117}$$

Moreover, an initial value for x_n, sufficiently close to the root in question, must be given.

For simplicity and generality, it will be assumed in the flow diagram that expressions for both $f(x)$ and $f'(x)$ are available. The approximate solution x_0 will also be presumed.

Figure 8.12 is a flow diagram for the determination of real roots of nonlinear algebraic equations using Newton's first-order method. The notes in the figure correspond to the following:

1. Computation begins with the introduction of the approximate value of the root in question, x_0. An allowable error in x, which governs the accuracy of the desired root, must also be specified.
2. A, the last calculated value of the root, is first set equal to x_0. $f(x)$ and $f'(x)$ at $x = A$ are then evaluated. In the flow diagram, FA has been used as the designation for $f(A)$. $F1A$ stands for $f'(A)$. It should be noted that if $f'(x) = 0$, the second term of Eq. (8.117) will go to infinity. Therefore, in the flow diagram, if $F1A = 0$, computation stops and a new value of x_0 must be selected.
3. B is the calculated "next approximation" to the root in question.
4. If the absolute value of the difference between B and A is found to be less than *Error*, an acceptable root has been obtained. If, however, the difference is too large, A is replaced by B and the process is repeated.

Prob. 8.10: Construct a flow diagram for the determination of the root of a nonlinear algebraic equation, using the method of false position.

Prob. 8.11: Develop a flow diagram for the preparation of a table of fifth roots, for numbers ranging from 0 to 50.0 in increments of 1.0, using Newton's first-order method. The table should be accurate to three decimal places.

Prob. 8.12: Using the method of solution described in Sec. 8.6.1, develop a flow diagram for the determination of the real roots of a set of simultaneous nonlinear algebraic equations.

miscellaneous methods

9

The first part of this text, Chaps. 1 to 8, is concerned primarily with the manipulation of numbers and groups of numbers and/or symbols, the emphasis being directed toward the solution of a given equation or set of equations. It is evident from a consideration of the Table of Contents that a major part of the second portion, Chaps. 10 and 11, is concerned with the development of methods for the numerical solution of differential equations. Many important numerical problems in engineering and applied science do not conveniently fit into these categories, for example, interpolation of tabulated data, curve fitting, and numerical integration. Chapter 9 is included to cover a few of these "other topics."

9.1 Difference Tables

To facilitate presentation and to provide a more compact form for future reference, scientific and mathematical data are frequently arranged in tabular form. For example, tables of Bessel functions are available which list, for equal increments of the argument, the appropriate function values. Sines, cosines, square roots, interest tables, and location of the position of the moon with respect to the sun and the earth are all data that are arranged in this form.

In dealing with tabular quantities, three types of situations are normally encountered. First, given the necessary mathematical relationship(s), it may be desired to develop a function table or to extend the range of one that has already been defined. Secondly, interpolation between function values is often required. The third case, referred to as inverse interpolation, has to do with the determination of that particular x which corresponds to a given $f(x)$, when the given $f(x)$ does not correspond identically to one of those tabulated.

These and other topics associated with the use of tabulated quantities can be more conveniently discussed after consideration of the development of a *difference table.* As is shown in a later chapter, such tables are also helpful in the solution of certain types of differential equations.

The specific equation that is used for illustration is the polynomial

$$\boxed{f(x) = 3x^2 + x + 2} \tag{9.1}$$

It is assumed that the range of interest is $3 \leq x \leq 8$ and that the increment, or interval, is 1. Without consideration of how they are obtained, the values of the function at the six points in question are as given in Table 9.1. By choosing as a positive system one that assumes that $f(x)$ increases with x, it is noted that the increment in the function values between $x = 3$ and $x = 4$ is

Table 9.1

x	+3	+4	+5	+6	+7	+8
$f(x)$	+32	+54	+82	+116	+156	+202

$$\Delta f_{(3\text{-}4)} = f(4) - f(3) = (+54) - (+32) = +22 \tag{9.2}$$

That is, the difference between the two quantities is +22. Similarly, the increment in the function values at $x = 4$ and $x = 5$ is

$$\Delta f_{(4\text{-}5)} = f(5) - f(4) = (+82) - (+54) = +28 \tag{9.3}$$

For the total range in question

$$\boxed{\begin{aligned} \Delta f_{(3\text{-}4)} &= +22 \\ \Delta f_{(4\text{-}5)} &= +28 \\ \Delta f_{(5\text{-}6)} &= +34 \\ \Delta f_{(6\text{-}7)} &= +40 \\ \Delta f_{(7\text{-}8)} &= +46 \end{aligned}} \tag{9.4}$$

It should be noted that these *first differences* are not constants. It is therefore possible to consider the differences (or increments) that exist between these computed values. That is, *second differences* can be obtained.

$$\Delta^2 f_{(3\text{-}4\text{-}5)} = (+28) - (+22) = +6 \tag{9.5}$$

where $\Delta^2 f$ is read as the second difference in the function value $f(x)$. Similarly,

$$\boxed{\begin{aligned} \Delta^2 f_{(4\text{-}5\text{-}6)} &= +6 \\ \Delta^2 f_{(5\text{-}6\text{-}7)} &= +6 \\ \Delta^2 f_{(6\text{-}7\text{-}8)} &= +6 \end{aligned}} \tag{9.6}$$

Extending the concept,

$$\boxed{\begin{aligned} \Delta^3 f_{(3\text{-}4\text{-}5\text{-}6)} &= 0 \\ \Delta^3 f_{(4\text{-}5\text{-}6\text{-}7)} &= 0 \\ \Delta^3 f_{(5\text{-}6\text{-}7\text{-}8)} &= 0 \end{aligned}} \tag{9.7}$$

While the development presented above is quite straightforward, a more convenient form for presentation of the resulting information is shown in Table 9.2. It is to be noted that at each stage the difference values are located in the difference table between the two primary values in question. That $\Delta^3 f = 0$ in Table 9.2 is not unexpected. Consider the same situation, but this time in terms of the general locations x, $x + 1$, $x + 2$, etc. (See Table 9.3.)

By noting that $\Delta^3 f \equiv 0$, Tables 9.2 and 9.3 can be extended indefinitely by simple addition, starting from the right-hand side and working to the left, and from the top downward.

Table 9.2

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
3	+32			
		+22		
4	+54		+6	
		+28		0
5	+82		+6	
		+34		0
6	+116		+6	
		+40		0
7	+156		+6	
		+46		
8	+202			

Table 9.3

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
x	$3x^2 + x + 2$			
		$+6x + 4$		
$x + 1$	$3x^2 + 7x + 6$		$+6$	
		$+6x + 10$		0
$x + 2$	$3x^2 + 13x + 16$		$+6$	
		$+6x + 16$		0
$x + 3$	$3x^2 + 19x + 32$		$+6$	
		$+6x + 22$		
$x + 4$	$3x^2 + 25x + 54$			

Table 9.4

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
3	+32			
		+22		
4	+54		+6	
		+28		0
5	+82		+6	
		+34		0
6	+116		+6	
		+40		0
7	+156		+6	
		+46		(0)
8	+202		(+6)	
		(+52)		(0)
(9)	(+254)		(+6)	
		(+58)		(0)
(10)	(+312)		(+6)	
		(+64)		(0)
(11)	(+376)		(+6)	
		(+70)		(0)
(12)	(+446)		(+6)	
		(+76)		(0)
(13)	(+522)		(+6)	
		(+82)		
(14)	(+604)			

Certain checks on the numerical work are readily apparent. For example, the first tabulated value of the first difference (that is, $\Delta f_{(3\text{-}4)}$) in Table 9.4 is +22. The total change in Δf over the range shown is $\Sigma(\Delta^2 f)$, or $(10)(+6) = +60$. Therefore the ending value in the Δf column should be $+22 + 60 = +82$. Similarly, $f(3) = +32$, $\Sigma(\Delta f) = +572$. Therefore, $f(14) = +32 + 572 = +604$.

Seldom are engineering data encountered in the convenient integral-number form of the preceding example. More often, decimal values are given and a certain specified accuracy is desired. Consider the same polynomial, evaluated in the range $3.00 \le x \le 4.50$. The increments are to be 0.25, and two decimal places are to be carried in the tabulation. (See Table 9.5.) As is customary in a difference table of this type, decimal points have been omitted.

Table 9.5

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
3.00	+32.00			
		+493		
3.25	+36.93		+39	
		+532		−3
3.50	+42.25		+36	
		+568		+3
3.75	+47.93		+39	
		+607		−3
4.00	+54.00		+36	
		+643		+3
4.25	+60.43		+39	
		+682		
4.50	+67.25			

DECIMAL POINTS OMITTED

EXTENDING

It should be noted that the third difference in Table 9.5 is not zero, as was the case in Table 9.4. Moreover, higher-order differences such as $\Delta^4 f$ and $\Delta^5 f$ are also not of value zero. Since $\Delta^3 f$ is constant in magnitude and alternating in sign, the difference table could be developed indefinitely, and nonzero values would continue to be obtained. This situation, however, does not rule out the possibility of "additive extension" of the table. In view of the repetitive, alternating nature of the values in the $\Delta^3 f$ column, the table can be extended, as shown in Table 9.6.

For the same polynomial [Eq. (9.1)], Table 9.7 is a difference table in the range $3.00 \le x \le 3.05$, where the uniform increment in x is 0.01.

By comparing Tables 9.4, 9.6, and 9.7, it is evident that the interval of the independent variable is of importance in determining the number of differences that must be considered to facilitate the extension of a function-value table. In general, higher-order differences must be determined until either a zero is obtained or the magnitudes of the increments are essentially equal and the signs are alternating. When such situations cannot be realized, the interval is too large and a smaller one must be selected.

IF STEADY STATE DOES NOT OCCUR

Table 9.6

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
4.00	+54.00		+36	
		+643		+3
4.25	+60.43		+39	
		+682		(−3)
4.50	+67.25		(+36)	
		(+718)		(+3)
(4.75)	(+74.43)		(+39)	
		(+757)		(−3)
(5.00)	(+82.00)		(+36)	
		(+793)		(+3)
(5.25)	(+89.93)		(+39)	
		(+832)		(−3)
(5.50)	(+98.25)		(+36)	
		(+868)		(+3)
(5.75)	(+106.93)		(+39)	
		(+907)		
(6.00)	(+116.00)			

Table 9.7

x	$f(x)$	Δf	$\Delta^2 f$
3.00	32.00		
		+19	
3.01	32.19		0
		+19	
3.02	32.38		0
		+19	
3.03	32.57		0
		+19	
3.04	32.76		0
		+19	
3.05	32.95		(0)
		(+19)	
(3.06)	(33.14)		(0)
		(+19)	
(3.07)	(33.33)		(0)
		(+19)	
(3.08)	(33.52)		(0)
		(+19)	
(3.09)	(33.71)		(0)
		(+19)	
(3.10)	(33.90)		

From the difference tables that have thus far been developed, it is evident that the first differences are determined from a consideration of the function values at two adjacent positions in the table. Second differences, on the other hand, are obtained by considering the differences between the first differences straddling the point in question, but these could also be written in terms of the function values themselves.

$$\Delta^2 f_{(1\text{-}2\text{-}3)} = \Delta f_{(2\text{-}3)} - \Delta f_{(1\text{-}2)} \tag{9.8}$$

or

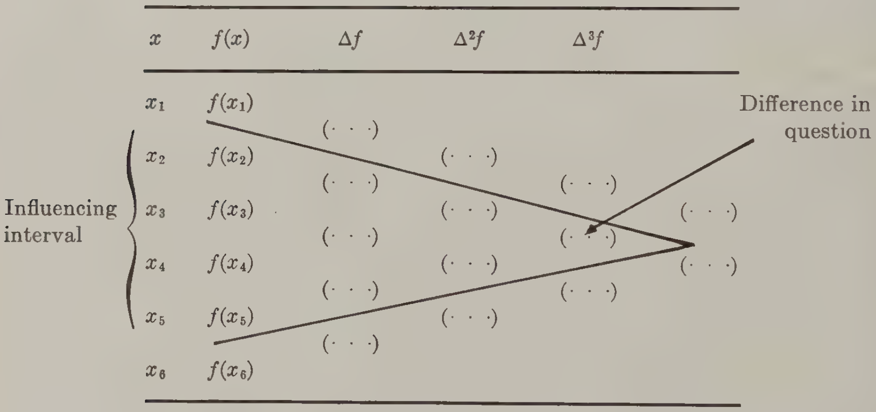
$$\Delta^2 f_{(1\text{-}2\text{-}3)} = (f_3 - f_2) - (f_2 - f_1)$$

or

$$\Delta^2 f_{(1\text{-}2\text{-}3)} = f_3 - 2f_2 + f_1 \tag{9.9}$$

Second differences are therefore determined from a consideration of three consecutive function values. In a similar manner, it can be shown that the third differences are dependent on four consecutive function values. This is shown diagrammatically by the triangular wedge of Table 9.8. Regardless of the order, with the apex of the wedge containing the appropriate difference in question, the range of influencing function values can be determined from a consideration of the type shown in the table.

Table 9.8



Not all data are available in equal increments of the independent variable. This is especially true when tabulations are prepared from experimental results. For such cases a difference table that takes into account the magnitude of the increment in x by dividing the difference in question by the appropriate influencing interval (see Table 9.8) is extremely useful. Such a table is called a *divided-difference table,* and the differences are noted symbolically as Δf, $\Delta^2 f$, etc. Table 9.9 is a divided-difference

Table 9.9

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
3	+32			
		+22		
4	+54		+3	
		+28		0
5	+82		+3	
		+37		0
7	+156		+3	
		+52		0
10	+312		+3	
		+64		0
11	+376		+3	
		+73		
13	+522			

table for the example given previously as Table 9.4. The first divided difference enclosed by the dashed oval is obtained as shown in Eq. (9.10).

$$\Delta f = \frac{(+312) - (+156)}{(+10) - (+7)} = \frac{+156}{+3} = +52 \tag{9.10}$$

Similarly, the designated second divided difference is

$$\Delta^2 f = \frac{(+52) - (+37)}{(+10) - (+5)} = \frac{+15}{+5} = +3 \tag{9.11}$$

Here, as in the preceding case, the second differences are constants and the table can be extended in a comparatively elementary fashion.

9.2 Errors in Difference Tables

If the error in question is an isolated misentry in a given function value, it is analytically possible to locate the error and to determine its value. Consider, for example, Table 9.10. The function in question is the same as that for which Table 9.4 was developed: however, for illustration, an error of +10 has been purposely introduced in entry $f(8)$. Large deviations from the values that were determined in Table 9.4 are noted in the region of the table enclosed within the dotted lines. Moreover, it is observed that as the order of the difference increases, the deviations from previously calculated values alternate in sign. This error is propagated in the manner that would be expected from a single error source ϵ, as demonstrated in Table 9.11. Only the error parts of the table are shown.

Table 9.10

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
3	+32				
		+22			
4	+54		+6		
		+28		0	
5	+82		+6		0
		+34		0	
6	+116		+6		+10
		+40		+10	
7	202 = +156		+16		−40
		+56		−30	
8	+212		−14		+60
		+42		+30	
9	202 = +254		+16		−40
		+58		−10	
10	+312		+6		+10
		+64		0	
11	+376		+6		0
		+70		0	
12	+446		+6		
		+76			
13	+522				

212
202
=10

12
202
=10

By comparing the fourth-difference columns of Tables 9.10 and 9.11, and assuming that the value of the fourth differences should be zero, the error can be determined from any of the following equations:

$$+\epsilon = +10$$

$$-4\epsilon = -40 \tag{9.12}$$

$$+6\epsilon = +60$$

The error-propagation table for two adjacent misentries, assuming that both errors are equal and of the same sign, is given as Table 9.12. Other error-propagation tables could be developed in a similar manner.

For the single-misentry case, illustrated in Tables 9.10 and 9.11, the largest deviations from expected values, for the even-numbered differences (that is, $\Delta^2 f$, $\Delta^4 f$, etc.), are contained in the same line as the initial-function-value error. This information is helpful in locating the more probable error sources in the given table. For two adjacent misentries, shown in Table 9.12, the location of the error source may not be as obvious. This is particularly true when entries have been rounded off.

ERROR
TABLE

Table 9.11

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_{i-3}	0		0		0
		0		0	
x_{i-2}	0		0		$+\epsilon$
		0		$+\epsilon$	
x_{i-1}	0		$+\epsilon$		-4ϵ
		$+\epsilon$		-3ϵ	
x_i	$+\epsilon$		-2ϵ		$+6\epsilon$
		$-\epsilon$		$+3\epsilon$	
x_{i+1}	0		$+\epsilon$		-4ϵ
		0		$-\epsilon$	
x_{i+2}	0		0		$+\epsilon$
		0		0	
x_{i+3}	0		0		0

Table 9.12

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_{i-3}	0		0		$+\epsilon$
		0		$+\epsilon$	
x_{i-2}	0		$+\epsilon$		-3ϵ
		$+\epsilon$		-2ϵ	
x_{i-1}	$+\epsilon$		$-\epsilon$		$+2\epsilon$
		0		0	
x_i	$+\epsilon$		$-\epsilon$		$+2\epsilon$
		$-\epsilon$		$+2\epsilon$	
x_{i+1}	0		$+\epsilon$		-3ϵ
		0		$-\epsilon$	
x_{i+2}	0		0		$+\epsilon$
		0		0	
x_{i+3}	0		0		0

9.3 The Difference Operators Δ, ∇, and δ

To be able to develop compact, analytical expressions for the various interrelationships that exist between the entries in a difference table, and to use these in the development of interpolation and extrapolation formulas, it is first necessary to define the symbolic linear operators Δ, ∇, and δ. By definition:

Δ = forward-difference operator
∇ = backward-difference operator
δ = central-difference operator

E = SHIFT OPERATOR

The shift operator, denoted by the letter E, is also of importance.

It should be understood at the outset that the introduction of these symbolic operators will not alter the magnitudes of the various entries in a given difference table. Their purpose is to facilitate understanding of the basic concepts and to provide a more usable form for the presentation of results.

In the work that will be discussed in this section, it has been assumed that the independent variable x is subdivided into *equally spaced intervals* of magnitude h. It has also been presumed that as x increases there will be a positive increase in the function $y = f(x)$ and all its differences $\Delta^n f$.

I

The *forward-difference operator* Δ assumes that the difference in question is referred to the least of the considered x positions. That is,

$$\boxed{\Delta^n f(x) = \Delta^{n-1} f(x+h) - \Delta^{n-1} f(x)} \tag{9.13}$$

For the first difference, referred to position 0,

$$\Delta f_0 = f_1 - f_0 \tag{9.14}$$

Similarly,

$$\Delta^2 f_0 = \Delta f_1 - \Delta f_0 = (f_2 - f_1) - (f_1 - f_0)$$

or

$$\Delta^2 f_0 = f_2 - 2f_1 + f_0 \tag{9.15}$$

In like manner,

$$\Delta^3 f_0 = \Delta^2 f_1 - \Delta^2 f_0 = (f_3 - 2f_2 + f_1) - (f_2 - 2f_1 + f_0)$$

or

$$\Delta^3 f_0 = f_3 - 3f_2 + 3f_1 - f_0 \tag{9.16}$$

The third forward difference referred to position 3 is given in Eq. (9.17).

$$\Delta^3 f_3 = \Delta^2 f_4 - \Delta^2 f_3 = f_6 - 3f_5 + 3f_4 - f_3 \tag{9.17}$$

In all these cases it is to be noted that the difference referred to a particular position contains the function values at that position and at those forward of it; thus the notation *forward differences* and *forward-difference operator*. A difference table in symbolic forward differences is given as Table 9.13. A forward-sloping line, to emphasize the forward differences referred to the 0 position, is shown in the table.

The *shift operator E* is used to define the value of the function at position $x + h$, that is, $f(x + h)$, in terms of $f(x)$;

$$\boxed{f(x+h) = Ef(x)} \tag{9.18}$$

Table 9.13

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
-3	y_{-3}				
		Δy_{-3}			
-2	y_{-2}		$\Delta^2 y_{-3}$		
		Δy_{-2}		$\Delta^3 y_{-3}$	
-1	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$
		Δy_{-1}		$\Delta^3 y_{-2}$	
0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$
		Δy_0		$\Delta^3 y_{-1}$	
$+1$	y_{+1}		$\Delta^2 y_0$		$\Delta^4 y_{-1}$
		Δy_{+1}		$\Delta^3 y_0$	
$+2$	y_{+2}		$\Delta^2 y_{+1}$		$\Delta^4 y_0$
		Δy_{+2}		$\Delta^3 y_{+1}$	
$+3$	y_{+3}		$\Delta^2 y_{+2}$		$\Delta^4 y_{+1}$

Since these function values are also related through forward differences, it should be possible to determine the relationship that exists between Δ and E. For illustration, consider the first difference referred to position x.

$$\Delta f_x = f_{x+h} - f_x = Ef_x - f_x = (E-1)f_x$$

The relation between the operators is therefore

$$\Delta = E - 1$$

or

$$E = 1 + \Delta \tag{9.19}$$

For second-order difference

$$\Delta^2 f_x = f_{x+2h} - 2f_{x+h} + f_x = (E-2)f_{x+h} + f_x$$

or

$$\Delta^2 f_x = (E-1)^2 f_x \tag{9.20}$$

By extending this concept, it can be shown that the forward-difference operator Δ and the shift operator E are related according to Eq. (9.21).

$$\Delta^n f_x = (E-1)^n f_x$$

$$\Delta^n = (E-1)^n \tag{9.21}$$

II

The *backward-difference operator* ∇ assumes that the difference in question is referred to the greater of the considered x positions. That is,

$$\boxed{\nabla^n f_x = \nabla^{n-1} f_{(x)} - \nabla^{n-1} f_{(x-h)}} \tag{9.22}$$

By noting from Table 9.8 the interval of function values which influences a particular order difference, it is evident that backward differences will

$$\nabla f_x = f_{(x)} - f_{(x-h)}$$

contain the function value of the point in question and the values of those points backward from it. A difference table in symbolic backward differences is given as Table 9.14. As in the preceding forward-difference

Table 9.14

x	$f(x)$	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
-3	y_{-3}		$\nabla^2 y_{-2}$		$\nabla^4 y_{-1}$
		∇y_{-2}		$\nabla^3 y_{-1}$	
-2	y_{-2}		$\nabla^2 y_{-1}$		$\nabla^4 y_0$
		∇y_{-1}		$\nabla^3 y_0$	
-1	y_{-1}		$\nabla^2 y_0$		$\nabla^4 y_{+1}$
		∇y_0		$\nabla^3 y_{+1}$	
0	y_0		$\nabla^2 y_{+1}$		$\nabla^4 y_{+2}$
		∇y_{+1}		$\nabla^3 y_{+2}$	
$+1$	y_{+1}		$\nabla^2 y_{+2}$		$\nabla^4 y_{+3}$
		∇y_{+2}		$\nabla^3 y_{+3}$	
$+2$	y_{+2}		$\nabla^2 y_{+3}$		
		∇y_{+3}			
$+3$	y_{+3}				

case, a sloping line has been shown to emphasize the reference locations. The 0 position was chosen for illustration.

By noting that

$$f(x) = Ef(x - h) \qquad \text{or} \qquad f(x - h) = \frac{1}{E} f(x) \tag{9.23}$$

the first backward difference can be written in terms of the shift operator.

$$\nabla f_x = f_x - f_{x-h} = f_x - \frac{1}{E} f_x$$

Therefore,

$$\Delta f_x = \left(1 - \frac{1}{E}\right) f_x = (1 - E^{-1}) f_x \tag{9.24}$$

or

$$\nabla = (1 - E^{-1}) \tag{9.25}$$

This, in turn, can be related to the forward-difference operator.

$$\nabla = \frac{\Delta}{1 + \Delta} \qquad \text{and} \qquad \Delta = \frac{\nabla}{1 - \nabla} \tag{9.26}$$

It would seem, from the two cases just considered, that the linear operators Δ, ∇, and E behave in the same manner as algebraic symbols. This, in fact, is the case, and it can be shown that all the operators Δ, ∇, δ, and E satisfy the distributive, associative, and commutative laws of algebra.

The *central-difference operator* δ is used when the reference locations are contained in a horizontal line in the difference table. A difference table in symbolic central differences is given as Table 9.15. With such a system, central-difference expressions will contain the same number of function values forward of the reference location in question as backward from it.

III

Table 9.15

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
-3	y_{-3}				
		$\delta_{-5/2}$			
-2	y_{-2}		$\delta_{-2}{}^2$		
		$\delta_{-3/2}$		$\delta_{-3/2}{}^3$	
-1	y_{-1}		$\delta_{-1}{}^2$		$\delta_{-1}{}^4$
		$\delta_{-1/2}$		$\delta_{-1/2}{}^3$	
0	y_0		$\delta_0{}^2$		$\delta_0{}^4$
		$\delta_{+1/2}$		$\delta_{+1/2}{}^3$	
$+1$	y_{+1}		$\delta_{+1}{}^2$		$\delta_{+1}{}^4$
		$\delta_{+3/2}$		$\delta_{+3/2}{}^3$	
$+2$	y_{+2}		$\delta_{+2}{}^2$		
		$\delta_{+5/2}$			
$+3$	y_{+3}				

In central differences

$$\delta f\left(x + \frac{h}{2}\right) = f(x + h) - f(x) = (E - 1)f(x)$$

or

$$\delta E^{1/2} f(x) = (E - 1)f(x) \tag{9.27}$$

This gives

$$\delta = E^{+1/2} - E^{-1/2} \tag{9.28}$$

By substituting the value of E from Eq. (9.19), δ can be expressed in terms of Δ.

$$\delta = \Delta(1 + \Delta)^{-1/2} \tag{9.29}$$

Another useful relationship can be established by multiplying both sides of Eq. (9.28) by $+E^{+1/2}$.

$$\delta E^{+1/2} = E - 1 = \Delta \tag{9.30}$$

By squaring both sides of this equation, it is seen that the second-difference operators δ^2 and Δ^2 are related through the shift operator E.

$$\delta^2 E = \Delta^2 \tag{9.31}$$

For example, the second central difference referred to location 0 equals the second forward difference referred to location -1.

$$\delta^2(Ey_{-1}) = \Delta^2 y_{-1}$$

or

$$\delta^2 y_0 = \Delta^2 y_{-1} \qquad (9.32)$$

A comparison of the appropriate entries in Tables 9.13 and 9.15 verifies this relationship.

Table 9.16 is a summary of several of the relationships that exist between the linear operators Δ, ∇, δ, and E.

Table 9.16

	$\Delta = \nabla(1 - \nabla)^{-1}$	$\Delta = E - 1$
$\nabla = \Delta(1 + \Delta)^{-1}$		$\nabla = (1 - E^{-1})$
$\delta = \Delta(1 + \Delta)^{-1/2}$	$\delta = \nabla(1 - \nabla)^{-1/2}$	$\delta = E^{+1/2} - E^{-1/2}$
$E = 1 + \Delta$	$E = (1 - \nabla)^{-1}$	

It is interesting to note that these linear operators can also be related to the operator $D = d/dx$ in differential equations. If it is assumed that the function $f(x)$ in question is single-valued and continuous in the range in question, expansion in a Taylor series about point x yields Eq. (9.33).

$$f(x + h) = f(x) + \frac{h}{1!} f'(x) + \frac{h^2}{2!} f''(x) + \cdots$$

or

$$Ef(x) = f(x + h) = \left(1 + \frac{hD}{1!} + \frac{h^2D^2}{2!} + \cdots\right) f(x) \qquad (9.33)$$

The term in the parentheses, however, represents the series e^{hD}. Therefore, in symbolic form

$$E = 1 + \frac{hD}{1!} + \frac{h^2D^2}{2!} + \cdots = e^{hD} \qquad (9.34)$$

Equation (9.34) is the fundamental linking relationship between numerical differencing and the formal analytical processes of differentiating and integration. The interrelationship can be extended to other linear operators considered in this section, which are shown *symbolically* in Eqs. (9.35).

$$\begin{aligned} E &= e^{hD} \\ \Delta &= e^{hD} - 1 \\ \nabla &= 1 - e^{-hD} \\ \delta &= 2 \sinh \tfrac{1}{2}hD \end{aligned} \qquad (9.35)$$

DERIVATIVE OF TABLE DATA

The operators Δ and ∇ can also be related, in expanded form, to the operator D. Given the basic symbolic relationship, repeated here as Eq. (9.36),

$$\Delta = e^{hD} - 1 \tag{9.36}$$

series expansion will yield

$$\Delta = hD + \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} + \frac{h^4D^4}{4!} + \cdots \tag{9.37}$$

The second forward difference could be obtained in a similar manner.

$$\Delta^2 = (e^{hD} - 1)^2 = e^{2hD} - 2e^{hD} + 1 = \cdots \tag{9.38}$$

Alternatively, by recognizing that linear operators can be treated the same as algebraic quantities, Δ^2 can be obtained directly from Eq. (9.37).

$$\Delta^2 = \left(hD + \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} + \frac{h^4D^4}{4!} + \cdots\right)^2$$
$$= h^2D^2 + h^3D^3 + \tfrac{7}{12}\, h^4D^4 + \cdots \tag{9.39}$$

Expressions can be obtained for higher-order forward differences as well as for backward differences in a similar manner. A summary of several of these is given in Table 9.17. The inverse system can also be defined by series expansion, and these are summarized in Table 9.18.

Table 9.17

$\Delta = hD + \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} + \frac{h^4D^4}{4!} + \cdots$	$\nabla = hD - \frac{h^2D^2}{2!} + \frac{h^3D^3}{3!} - \frac{h^4D^4}{4!} + \cdots$
$\Delta^2 = h^2D^2 + h^3D^3 + \tfrac{7}{12}h^4D^4 + \cdots$	$\nabla^2 = h^2D^2 - h^3D^3 + \tfrac{7}{12}h^4D^4 - \cdots$
$\Delta^3 = h^3D^3 + \tfrac{3}{2}h^4D^4 + \tfrac{5}{4}h^5D^5 + \cdots$	$\nabla^3 = h^3D^3 - \tfrac{3}{2}h^4D^4 + \tfrac{5}{4}h^5D^5 - \cdots$

Table 9.18

FORWARD	BACKWARD
$D = \frac{1}{h}\left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \frac{\Delta^4}{4} + \cdots\right)$	$D = \frac{1}{h}\left(\nabla + \frac{\nabla^2}{2} + \frac{\nabla^3}{3} + \frac{\nabla^4}{4} + \cdots\right)$ Δ_{-1} Δ_{-2} Δ_{-3}
$D^2 = \frac{1}{h^2}\left(\Delta^2 - \Delta^3 + \tfrac{11}{12}\Delta^4 - \tfrac{5}{6}\Delta^5 + \cdots\right)$	$D^2 = \frac{1}{h^2}\left(\nabla^2 + \nabla^3 + \tfrac{11}{12}\nabla^4 + \tfrac{5}{6}\nabla^5 + \cdots\right)$
$D^3 = \frac{1}{h^3}\left(\Delta^3 - \tfrac{3}{2}\Delta^4 + \tfrac{7}{4}\Delta^5 - \cdots\right)$	$D^3 = \frac{1}{h^3}\left(\nabla^3 + \tfrac{3}{2}\nabla^4 + \tfrac{7}{4}\nabla^5 + \cdots\right)$
$D^4 = \frac{1}{h^4}\left(\Delta^4 - 2\Delta^5 + \tfrac{17}{6}\Delta^6 - \cdots\right)$	$D^4 = \frac{1}{h^4}\left(\nabla^4 + 2\nabla^5 + \tfrac{17}{6}\nabla^6 + \cdots\right)$

GOOD FOR ANY CASE WHEN h IS CONSTANT

It should be understood that differential or difference operators of the type given in Table 9.18 can equally well be listed in terms of the function values themselves. For example, from Eqs. (9.14) to (9.16),

$$\begin{aligned} \Delta f_0 &= f_1 - f_0 \\ \Delta^2 f_0 &= f_2 - 2f_1 + f_0 \\ \Delta^3 f_0 &= f_3 - 3f_2 + 3f_1 - f_0 \end{aligned} \tag{9.40}$$

Therefore, since

$$D = \frac{1}{h}\left(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \cdots\right)$$

$$Df_0 = \frac{1}{h}\left(\Delta f_0 - \tfrac{1}{2}\Delta^2 f_0 + \tfrac{1}{3}\Delta^3 f_0 - \cdots\right)$$

or $$Df_0 = \frac{1}{h}\left[(f_1 - f_0) - \tfrac{1}{2}(f_2 - 2f_1 + f_0) + \tfrac{1}{3}(f_3 - 3f_2 + 3f_1 - f_0) - \cdots\right]$$

which gives

$$Df_0 = \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3 - \cdots) \tag{9.41}$$

9.4 Interpolation

As pointed out in Sec. 9.1, one of the most often encountered problems associated with the use of tabulated quantities is that of interpolation. This is especially true when the function values are dependent on more than one variable. In this section, however, attention is directed toward the problem of interpolation in single-independent-variable systems, even though certain of the methods could equally well be extended to the more involved cases.

The basic problem in question is the development of numerical methods for calculating values of a function corresponding to values of the argument that lie between tabulated points. The most elementary form of such interpolation is referred to as *linear interpolation* or *taking proportional parts*. This assumes that in the range in question the function can be adequately represented by a linear function. As demonstrated later in this section, this is but a special, limiting case of the more general interpolation problem.

There are many approaches to the development of interpolation formulas, and several of these are considered in this section. However, consider first the "difference-table behavior" of certain known analytical relationships.

1. *y is a linear function in x.*

$$y = a_1x + a_0 \tag{9.42}$$

If the differencing interval is assumed to be h, the first forward difference is that given by Eq. (9.43).

$$\Delta y = f(x + h) - f(x) = (a_1)(x + h) + a_0 + (a_1x - a_0) = a_1h \tag{9.43}$$

Noting that the second and higher derivatives of y with respect to x are zero makes it possible to obtain this same relationship from the equations of Table 9.17:

$$\Delta y = hD(y) + 0 = h\frac{d}{dx}(a_1x + a_0) = ha_1 \tag{9.44}$$

Equation (9.45) is also evident.

$$\Delta^n y = 0 \qquad \text{for } n \geq 2 \tag{9.45}$$

2. *y is a quadratic in x.*

$$y = a_2x^2 + a_1x + a_0 \tag{9.46}$$

Using the equations of Table 9.17,

$$\begin{aligned} \Delta y &= h^2(a_2) + h(2a_2x + a_1) \\ \Delta^2 y &= h^2(2a_2) \\ \Delta^n y &= 0 \qquad \text{for } n \geq 3 \end{aligned} \tag{9.47}$$

3. *y is a cubic in x.*

$$y = a_3x^3 + a_2x^2 + a_1x + a_0 \tag{9.48}$$

$$\begin{aligned} \Delta y &= h^3(a_3) + h^2(3a_3x + a_2) + h(3a_3x^2 + 2a_2x + a_1) \\ \Delta^2 y &= h^3(6a_3) + h^2(6a_3x + 2a_2) \\ \Delta^3 y &= h^3(6a_3) \\ \Delta^n y &= 0 \qquad \text{for } n \geq 4 \end{aligned} \tag{9.49}$$

From these three examples and Table 9.17, it is evident that the nth difference of a polynomial of order n is a constant. Moreover, the value of that difference is

$$\boxed{\Delta^n y = h^n D^n y = a_n h^n n! = \nabla^n y} \tag{9.50}$$

The values of the $(n + 1)$st and all higher differences are zero.

$$D^n y = a_n (n!)$$

In view of the above, if the nth difference in a "regular" difference table is a constant, it is reasonable to assume that an nth-order polynomial can be made to represent the function in question, and this is one of the approaches used in developing interpolation formulas. It is to be noted, however, that there are certain analytic functions for which this assumption is not true. For example, should there be contained within the interval in question singularities or severe fluctuations, matching of the interpolating polynomial and the function at equally spaced "pivotal points" is not sufficient.

Figure 9.1 illustrates the general interpolation problem in question and defines the notation that is used. It is assumed that the function values are known at equally spaced increments of the argument and that a difference table either is available or can be defined. The problem is to determine the function value at $x + \theta h$. It is presumed that $\boxed{0 \leq \theta \leq 1.0.}$

Interpolation formulas are first developed, using the operator notation of Sec. 9.3. In terms of forward differences, the function value at $x + \theta h$ is

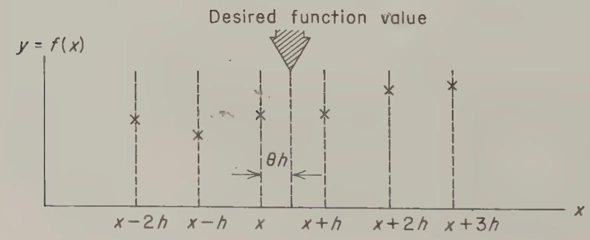
$$f(x + \theta h) = E^{\theta} f(x) = (1 + \Delta)^{\theta} f(x) \tag{9.51}$$

Expanding the term in parentheses by the binomial theorem yields *Newton's forward-difference interpolation formula:*

$$f(x + \theta h) = \left[1 + \theta \Delta + \frac{\theta(\theta - 1)}{2!} \Delta^2 + \cdots \right.$$
$$\left. + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - n + 1)}{n!} \Delta^n \right] f(x) \tag{9.52}$$

To illustrate the use of Eq. (9.52), assume that difference table 9.19 is given and that it is desired to evaluate the function at $x = 6.5$. (Note that Table 9.19 is a part of difference table 9.4.) Again, for emphasis a forward-sloping reference line has been drawn. To the right has also been included a symbolic forward-difference table.

Figure 9.1



x | y

Table 9.19

28 +8

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
		34	+6					
+6	+116				$f(6)$			
		+40				Δ_6		
+7	+156		+6		$f(7)$		Δ_6^2	
		+46		0		Δ_7		Δ_6^3
+8	+202		+6		$f(8)$		Δ_7^2	
		+52		0		Δ_8		Δ_7^3
+9	+254		+6		$f(9)$		Δ_8^2	
		+58				Δ_9		
+10	+312				$f(10)$			

For the case considered, $h = 1.0$ and the θ to be used in Eq. (9.52) is 0.5. Evaluation then proceeds as follows:

$$\begin{aligned} f(6.5) &= f(6.0 + \theta h) = f[6.0 + (0.5)(1.0)] \\ &= \left[1 + \theta\Delta + \frac{\theta(\theta - 1)}{2!}\Delta^2 + (0)\right] f(6) \\ &= f(6) + \theta\Delta_6 + \frac{\theta(\theta - 1)}{2!}\Delta_6^2 \\ &= +116 + (0.5)(+40) + \frac{(0.5)(0.5 - 1.0)}{2}(+6) \\ &= +135.25 \end{aligned} \tag{9.53}$$

The analytical function corresponding to Table 9.19 is $f(x) = 3x^2 + x + 2$. Substituting $x = 6.5$ in this equation results in the same value for $f(6.5)$ as that given by Eq. (9.53).

In the just preceding example the value of the argument, that is, $x = 6.5$, was near the beginning of the difference table, and consequently all the forward differences referred to $x = 6$ were defined. Consider now the situation where evaluation is desired for $x = 9.5$ and it is assumed that only the difference table given in 9.19 is available. Referring to the symbolic forward-difference table of 9.19 reveals that *only* $f(9)$ and Δ_9 are available for use in Eq. (9.52). For this and other cases where values are desired near the end of a difference table, *Newton's backward-difference interpolation formula* [Eq. (9.56)] should be used. From Table 9.16,

$$E = (1 - \nabla)^{-1} \tag{9.54}$$

Therefore,

$$f(x + \theta h) = E^\theta f(x) = (1 - \nabla)^{-\theta} f(x) \tag{9.55}$$

or

$$f(x + \theta h) = \left[1 + \theta\nabla + \frac{\theta(\theta + 1)}{2!}\nabla^2 + \cdots + \frac{\theta(\theta + 1)(\theta + 2)\cdots(\theta + n - 1)}{n!}\nabla^n\right]f(x) \qquad (9.56)$$

For $x = 9.5$, the reference point is $x = 10.0$ and $\theta = -0.5$. This gives

$$\begin{aligned} f(9.5) &= f(10.0 - 0.5) \\ &= f(10) + (-0.5)\nabla_{10} + \frac{(-0.5)(-0.5 + 1.0)}{2}\nabla_{10}^2 \\ &= +312 - (0.5)(+58) + \frac{(-0.5)(+0.5)}{2}(+6) = +282.25 \qquad (9.57) \end{aligned}$$

If it is known that the difference table is "regular," the reference point can be selected as $x = 9$ and $\theta = +0.5$ with the following result:

$$\begin{aligned} f(9.5) &= f(9.0 + 0.5) \\ &= +254 + (0.5)(+52) + \frac{(0.5)(0.5 + 1.0)}{2}(+6) = +282.25 \end{aligned} \qquad (9.58)$$

Even though this procedure results in exactly the same function value, it is to be recognized that this second case extrapolates beyond the range of influence of the differences used. It is therefore more subject to error and is not recommended for general use.

Equations (9.52) and (9.56) are but two of the many interpolation formulas that can be and have been developed from a consideration of the relationships listed in Table 9.16. Table 9.20 is a summary of several of these. For completeness, Eqs. (9.52) and (9.56) have also been included.

To illustrate the use of Gauss', Everett's, and Bessel's central-difference interpolation formulas, consider the evaluation of $f(8.5)$ for the difference table 9.19.

1. Gauss' forward-interpolation formula:

$$\begin{aligned} f(8.5) &= f(8.0 + 0.5) \\ &= +202 + (0.5)(+52) + \frac{(0.5)(-0.5)}{2}(+6) + 0 = +227.25 \end{aligned} \qquad (9.59a)$$

Θ = AMOUNT OFF OF THE INCREM
VALUE. $0 \leq \Theta \leq 1.0$

Table 9.20

h = INCREMENT

Newton's Forward-difference Interpolation Formula

$$f(x + \theta h) = f(x) + \theta\, \Delta f(x) + \frac{\theta(\theta - 1)}{2!} \Delta^2 f(x) + \cdots + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - n + 1)}{n!} \Delta^n f(x)$$

Newton's Backward-difference Interpolation Formula

$$f(x + \theta h) = f(x) + \theta\, \nabla f(x) + \frac{\theta(\theta + 1)}{2!} \nabla^2 f(x) + \cdots + \frac{\theta(\theta + 1)(\theta + 2) \cdots (\theta + n - 1)}{n!} \nabla^n f(x)$$

VALU
NE
EN
OF
TAB

∇y $\Delta^2 y$

Gauss' Forward-interpolation Formula (Illustrated for Location O)

$2r + 1 = n$

$\boxed{r = \frac{n-1}{2}}$

$$f(x + \theta h) = y_0 + \theta \overset{\Delta y_0}{(\delta y_{+\frac{1}{2}})} + \frac{\theta(\theta - 1)}{2!} \overset{\Delta^2 y_{-1}}{(\delta^2 y_0)} + \frac{\theta(\theta^2 - 1^2)}{3!} \overset{\Delta^3 y_{-1}}{(\delta^3 y_{+\frac{1}{2}})}$$
$$+ \frac{\theta(\theta^2 - 1^2)(\theta - 2)}{4!} \underset{\Delta^4 y_{-2}}{(\delta^4 y_0)} + \cdots + \frac{\theta(\theta^2 - 1^2) \cdots [\theta^2 - (r - 1)^2](\theta - r)}{(2r)!} \delta^{2r} y_0$$
$$+ \frac{\theta(\theta^2 - 1^2) \cdots (\theta^2 - r^2)}{(2r + 1)!} \delta^{2r+1} y_{+\frac{1}{2}}$$

Gauss' Backward-interpolation Formula (Illustrated for Location O)

$$f(x + \theta h) = y_0 + \theta \overset{\Delta y_{-1}}{(\delta y_{-\frac{1}{2}})} + \frac{\theta(\theta + 1)}{2!} \overset{\Delta^2 y_{-1}}{(\delta^2 y_0)} + \frac{\theta(\theta^2 - 1^2)}{3!} \overset{\Delta^3 y_{-2}}{(\delta^3 y_{-\frac{1}{2}})}$$
$$+ \frac{\theta(\theta^2 - 1^2)(\theta + 2)}{4!} \delta^4 y_0 + \cdots + \frac{\theta(\theta^2 - 1^2) \cdots [\theta^2 - (r - 1)^2](\theta + n)}{(2r)!} \delta^{2r} y_0$$
$$+ \frac{\theta(\theta^2 - 1^2) \cdots (\theta^2 - r^2)}{(2r + 1)!} \delta^{2r+1} y_{-\frac{1}{2}}$$

$r =$

Everett's Interpolation Formula (Illustrated for Location O)

$$f(x + \theta h) = \left[\theta y_{+1} + \frac{\theta(\theta - 1)(\theta + 1)}{3!} \delta^2 y_{+1} + \frac{\theta(\theta - 1)(\theta - 2)(\theta + 1)(\theta + 2)}{5!} \delta^4 y_{+1} + \cdots + \frac{\theta(\theta - 1) \cdots (\theta - r)(\theta + 1)(\theta + 2) \cdots (\theta + r - 1)}{(2r + 1)!} \delta^{2r} y_{+1} \right]$$
$$- \left[(\theta - 1) y_0 + \frac{\theta(\theta - 1)(\theta - 2)}{3!} \delta^2 y_0 + \frac{\theta(\theta - 1)(\theta - 2)(\theta - 3)(\theta + 1)}{5!} \delta^4 y_0 + \cdots + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - r - 1)(\theta + 1) \cdots (\theta + r - 1)}{(2r + 1)!} \delta^{2r} y_0 \right]$$

Table 9.20 (Continued)

Bessel's Interpolation Formula (Illustrated for Location O)

$$f(x+\theta h) = \left\{ \frac{y_0+y_1}{2} + \frac{\theta(\theta-1)}{2!}\left(\frac{\delta^2 y_0+\delta^2 y_1}{2}\right) + \frac{\theta(\theta-1)(\theta-2)(\theta+1)}{4!}\left(\frac{\delta^4 y_0+\delta^4 y_1}{2}\right) + \cdots + \frac{\theta(\theta^2-1^2)\cdots[\theta^2-(r-1)^2](\theta-r)}{(2r)!}\left(\frac{\delta^{2r} y_0+\delta^{2r} y_1}{2}\right)\right\}$$
$$+ \left\{ (\theta-\tfrac{1}{2})\,\delta y_{\frac{1}{2}} + \frac{\theta(\theta-1)(\theta-\frac{1}{2})}{3!}\,\delta^3 y_{\frac{1}{2}} + \frac{\theta(\theta-1)(\theta-2)(\theta-\frac{1}{2})(\theta+1)}{5!}\,\delta^5 y_{\frac{1}{2}} + \cdots + \frac{\theta(\theta^2-1^2)\cdots[\theta^2-(r-1)^2](\theta-r)(\theta-\frac{1}{2})}{(2r+1)!}\,\delta^{2r+1} y_{\frac{1}{2}} \right\}$$

2. Gauss' backward-interpolation formula:

$$f(8.5) = f(8.0+0.5) = +202 + (0.5)(+46) + \frac{(0.5)(1.5)}{2}(+6) + 0 = +227.25 \tag{9.59b}$$

3. Everett's interpolation formula:

$$f(8.5) = f(8.0+0.5) = \left[(0.5)(+254) + \frac{(0.5)(-0.5)(1.5)}{6}(+6)\right] - \left[(-0.5)(+202) + \frac{(0.5)(-0.5)(-1.5)}{6}(+6)\right] = +227.25 \tag{9.59c}$$

4. Bessel's interpolation formula:

$$f(8.5) = f(8.0+0.5) = \left[\frac{+202+254}{2} + \frac{(0.5)(-0.5)}{2}\,\frac{+6+6}{2}\right] + [(0)(+52)+0] = +227.25 \tag{9.59d}$$

With regard to the choice of a particular interpolation formula, the central-difference formulas will, in general, result in faster convergence. Of those listed, Bessel's and Everett's are recommended. However, when values near the end of a table are required, Newton's formulas must be used.

If the function represented by the difference table is a polynomial of degree n and the tabulated values are exact, the $(n + 1)$st and higher differences are zero, the interpolation formulas terminate, and interpolation is "exact." For those other cases where the $(n + 1)$st difference oscillates about zero, the function numerically responds as if it were a polynomial of degree n. The more "direct" development of polynomial interpolation formulas to represent tabulated function values would therefore seem in order.

POINT
MATC

Consider, for example, the situation defined in Fig. 9.2. It is presumed that three equally spaced points -1, 0, and $+1$ have associated with them the three function values y_{-1}, y_0, and y_{+1}. The origin is assumed to be located at position 0. It is desired to pass a single second-order polynomial of the type given by Eq. (9.60) through these three points.

$$\boxed{y = a_2x^2 + a_1x + a_0} \tag{9.60}$$

The three conditions necessary for a solution are given by Eq. (9.61).

SUBSTITUTING
y_{-1}, y_0, y_{+1}
$-1, 0, +1$

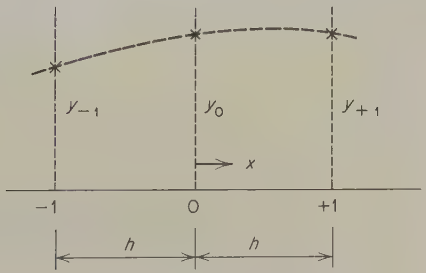
$$\left\{\begin{aligned} y_{-1} &= a_2(-h)^2 + a_1(-h) + a_0(1) \\ y_0 &= a_2(0) \quad\;\; + a_1(0) \quad\;\; + a_0(1) \\ y_{+1} &= a_2(+h)^2 + a_1(+h) + a_0(1) \end{aligned}\right. \tag{9.61}$$

These can be recast in the matrix form.

$$\begin{bmatrix} +h^2 & -h & +1 \\ 0 & 0 & +1 \\ +h^2 & +h & +1 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} y_{-1} \\ y_0 \\ y_{+1} \end{bmatrix} \qquad \text{or} \qquad [A][a] = [y] \tag{9.62}$$

The unknown coefficients a_i can be defined in terms of the known function

Figure 9.2



values y_i by calculating the inverse of $[A]$. That is,

$$[a] = [A^{-1}][y] \qquad \text{or} \qquad \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} +\dfrac{1}{2h^2} & -\dfrac{1}{h^2} & +\dfrac{1}{2h^2} \\ -\dfrac{1}{2h} & 0 & +\dfrac{1}{2h} \\ 0 & +1 & 0 \end{bmatrix} \begin{bmatrix} y_{-1} \\ y_0 \\ y_{+1} \end{bmatrix} \tag{9.63}$$

OINT-
ATCH

The corresponding central second-order interpolating polynomial is therefore

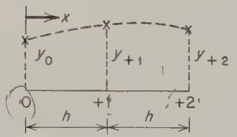
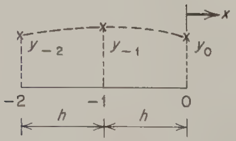
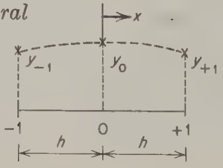
$$y = \frac{y_{-1} - 2y_0 + y_{+1}}{2h^2} x^2 + \frac{-y_{-1} + y_{+1}}{2h} x + y_0 \tag{9.64}$$

It is also possible to develop similar interpolating polynomials in terms of known function values forward of the point in question or backward from it. These are summarized in Table 9.21. They could equally well be written in terms of the notation used in Table 9.20, that is, $x/h = \theta$. Consider, for example, the second-order forward interpolating polynomial.

$$y = \frac{y_0 - 2y_{+1} + y_{+2}}{2h^2} x^2 + \frac{-3y_0 + 4y_{+1} - y_{+2}}{2h} x + y_0$$

$$y = \frac{y_0 - 2y_{+1} + y_{+2}}{2} \theta^2 + \frac{-3y_0 + 4y_{+1} - y_{+2}}{2} \theta + y_0$$

Table 9.21

Forward  x y_0 y_{+1} y_{+2} 0 +1 +2 h h	$y = \dfrac{y_0 - 2y_{+1} + y_{+2}}{2h^2} x^2 + \dfrac{-3y_0 + 4y_{+1} - y_{+2}}{2h} x + y_0$
Backward  x y_{-2} y_{-1} y_0 −2 −1 0 h h	$y = \dfrac{y_{-2} - 2y_{-1} + y_0}{2h^2} x^2 + \dfrac{y_{-2} - 4y_{-1} + 3y_0}{2h} x + y_0$
Central  x y_{-1} y_0 y_{+1} −1 0 +1 h h	$y = \dfrac{y_{-1} - 2y_0 + y_{+1}}{2h^2} x^2 + \dfrac{-y_{-1} + y_{+1}}{2h} x + y_0$

This can be rearranged as follows:

$$y = \frac{(y_0 - y_{+1}) - (y_{+1} - y_{+2})}{2}\theta^2 + \frac{-(y_0 - 2y_{+1} + y_{+2}) + (-2y_0 + 2y_{+1})}{2}\theta + y_0$$

$$= \frac{\Delta^2 y_0}{2}\theta^2 - \frac{\Delta^2 y_0}{2}\theta + \Delta y_0 \theta + y_0$$

or

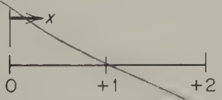
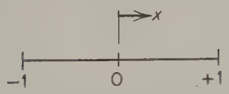
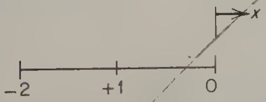
$$y = y_0 + \theta\,\Delta y_0 + \frac{\theta(\theta - 1)}{2}\Delta^2 y_0 \tag{9.65}$$

which corresponds to a terminated Newton's forward-difference interpolation formula. In similar manner, the central interpolating polynomial of Table 9.21 reduces to a terminated Gauss' (central-difference) backward-interpolation formula.

While the arrangement of the various terms in the interpolation equations of Table 9.21 is logical and consistent with the assumed polynomial (9.60), other forms could also be written. For example, in Table 9.22 each of the bracketed terms is a separate polynomial in x. Collection is with respect to the assumed known function values y_{-2}, y_{-1}, y_0, etc. The equations given in Table 9.22 suggest an interpolating polynomial made up of linear combinations of $n + 1$ separate polynomials in x. That is,

$$y = \sum_{i=0}^{n} L_i(x) y_i \tag{9.66}$$

Table 9.22

Forward  	$y = \left[\frac{(x - h)(x - 2h)}{(h)(2h)}\right] y_0 + \left[\frac{(x)(x - 2h)}{(h)(-h)}\right] y_{+1} + \left[\frac{(x)(x - h)}{(h)(2h)}\right] y_{+2}$
Central  	$y = \left[\frac{(x)(x - h)}{(h)(2h)}\right] y_{-1} + \left[\frac{(x - h)(x + h)}{(h)(-h)}\right] y_0 + \left[\frac{(x)(x + h)}{(h)(2h)}\right] y_{+1}$
Backward  	$y = \left[\frac{(x)(x + h)}{(h)(2h)}\right] y_{-2} + \left[\frac{(x)(x + 2h)}{(h)(-h)}\right] y_{-1} + \left[\frac{(x + h)(x + 2h)}{(h)(2h)}\right] y_0$

where the $[L_i(x)]$'s are nth-degree polynomials in the independent variable x and the summation includes a total of $n + 1$ terms. It is to be noted that these need not be referred specifically to the designated points $0, 1, 2, \ldots, n$. They could apply equally to the points $-3, -2, -1, 0, +1, \ldots, n - 3$, or any other $n + 1$ consecutive points.

In developing expressions for interpolating polynomials of the type given in Eq. (9.66), the more general approach of Lagrange is here used. That is, it will be presumed that *the function values are not necessarily given at equal spacing of the independent variable* x. Moreover, it is assumed that the individual polynomials are identically zero at all pivotal points other than i. For that one particular location, they equal $+1$. These conditions are defined by

$$[L_i(x_j)] = \begin{cases} 0 & \text{for } j \neq i \\ +1 & \text{for } j = i \end{cases} \tag{9.67}$$

From the first of these, by recognizing that each zero represents a root of the polynomial in question, Eq. (9.68) can be written immediately:

$$L_i(x_j) = A_j(x - x_0)(x - x_1) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n) \tag{9.68}$$

[It should be noted that the term $(x - x_j)$ is purposely missing from (9.68).] Applying the second of Eqs. (9.67),

$$1 = A_j(x_j - x_0)(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)$$

$$\text{or} \quad A_j = \frac{1}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)} \tag{9.69}$$

The individual polynomials are therefore defined by

$$L_i(x_j) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)} \tag{9.70}$$

In expanded form, these are

$$\begin{aligned} L_0(x) &= \frac{(x - x_1)(x - x_2) \cdots (x - x_n)}{(x_0 - x_1)(x_0 - x_2) \cdots (x_0 - x_n)} \\ L_1(x) &= \frac{(x - x_0)(x - x_2) \cdots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \cdots (x_1 - x_n)} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ L_n(x) &= \frac{(x - x_0)(x - x_1) \cdots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})} \end{aligned} \tag{9.71}$$

Lagrange's interpolation formula is frequently written in the abbreviated notation of Eq. (9.72).

$$y = \sum_{i=0}^{n} \prod_{\substack{j=0 \\ i \neq j}}^{n} \frac{x - x_j}{x_i - x_j} y_i \tag{9.72}$$

The expanded form is given as Eq. (9.73).

$$\begin{aligned} y = {} & \frac{(x - x_1)(x - x_2) \cdots (x - x_n)}{(x_0 - x_1)(x_0 - x_2) \cdots (x_0 - x_n)} y_0 \\ & + \frac{(x - x_0)(x - x_2) \cdots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \cdots (x_1 - x_n)} y_1 \\ & + \frac{(x - x_0)(x - x_1)(x - x_3) \cdots (x - x_n)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3) \cdots (x_2 - x_n)} y_2 + \cdots \\ & + \frac{(x - x_0)(x - x_1) \cdots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})} y_n \end{aligned} \tag{9.73}$$

If Eq. (9.73) is applied to the equal-spacing second-order central-difference case of Table 9.22, the interpolating formula will be developed as follows, starting with the transformations given in Table 9.23:

Table 9.23

Notation of Equation (9.70)	Second-order central-difference notation
x_0	x_{-1} (where $x_{-1} = -h$)
x_1	x_0 (where $x_0 = 0$)
x_2	x_{+1} (where $x_{+1} = +h$)

$$\begin{aligned} y = {} & \frac{(x - x_0)(x - x_1)}{(x_{-1} - x_0)(x_{-1} - x_1)} y_{-1} + \frac{(x - x_{-1})(x - x_1)}{(x_0 - x_{-1})(x_0 - x_1)} y_0 \\ & + \frac{(x - x_{-1})(x - x_0)}{(x_1 - x_{-1})(x_1 - x_0)} y_{+1} \\ = {} & \frac{(x - 0)(x - h)}{(-h - 0)(-h - h)} y_{-1} + \frac{(x + h)(x - h)}{(0 + h)(0 - h)} y_0 \\ & + \frac{(x + h)(x - 0)}{(h + h)(h - 0)} y_{+1} \end{aligned}$$

or

$$y = \frac{(x)(x - h)}{(h)(2h)} y_{-1} + \frac{(x - h)(x + h)}{(h)(-h)} y_0 + \frac{(x)(x + h)}{(h)(2h)} y_{+1} \tag{9.74}$$

It is to be noted that this is the same equation listed in Table 9.22. It should also be observed that in the development of Eqs. (9.71) the coefficients of the y's depend not on the specific values of x chosen, but only on the interval of spacing. The corresponding fourth-order central-difference interpolating polynomial is given as Eq. (9.75).

$$
\begin{aligned}
y = {} & \frac{(x)(x-h)(x-2h)(x+h)}{24h^4} y_{-2} + \frac{(x)(x-h)(x-2h)(x+2h)}{-6h^4} y_{-1} \\
& + \frac{(x-h)(x-2h)(x+h)(x+2h)}{4h^4} y_0 + \frac{(x)(x-2h)(x+h)(x+2h)}{-6h^4} y_{+1} \\
& + \frac{(x)(x-h)(x+h)(x+2h)}{24h^4} y_{+2} \qquad (9.75)
\end{aligned}
$$

To illustrate the use of the general Lagrange interpolation formula [Eq. (9.73)], consider the evaluation of $y = e^x$ for $x = 0.88372$, given the tabulated function values of Table 9.24. It is to be noted that only three values have been given, and these are not equally spaced. The interpolation equation corresponding to Eq. (9.73) is then

Table 9.24

x	$y = e^x$
0.85	2.3396
0.88	2.4109
0.90	2.4596

$$
\begin{aligned}
y = {} & \frac{(0.88372 - 0.88000)(0.88372 - 0.90000)}{(0.85 - 0.88)(0.85 - 0.90)} (2.3396) \\
& + \frac{(0.88372 - 0.85000)(0.88372 - 0.90000)}{(0.88 - 0.85)(0.88 - 0.90)} (2.4109) \\
& + \frac{(0.88372 - 0.85000)(0.88372 - 0.88000)}{(0.90 - 0.85)(0.90 - 0.88)} (2.4596) \qquad (9.76)
\end{aligned}
$$

or

$$
\begin{aligned}
y = {} & -(0.04037)(2.3396) + (0.91493)(2.4109) + (0.12544)(2.4596) \\
& = +2.4199 \qquad (9.77)
\end{aligned}
$$

To the four decimal places given, the answer is correct.

One of the advantages of using the Lagrange interpolation equation is that a difference table need not be constructed. Less computer storage is therefore required. It is to be noted, however, that as the procedure requires the subtraction of almost equal numbers, a large number of significant figures must be carried in the calculation to obtain the desired accuracy. A second shortcoming is that it is impossible to know at the outset the number of points needed for a particular accuracy. When points are added to improve the accuracy, all polynomials must be recalculated.

With regard to the general question of accuracy, *errors* in interpolation result from several sources. *Truncation errors,* for example, occur when terminated rather than the infinite series are used. The errors associated with such procedures correspond to the sums of the remaining terms of the infinite series. It is to be noted, however, that no general rule can be formulated for the definition of these remainders unless further knowledge of the tabulated functions is available. For that particular case where the nth difference terms are included in the interpolation, the $(n + 1)$st differences are constants, and the $(n + 2)$nd and higher differences are zero, the remainder can be defined. For Newton's forward-difference interpolation formula it is given by

$$R = \frac{h^{n+1}f^{[n+1]}(\xi)}{(n+1)!}[\theta(\theta - 1)(\theta - 2) \cdots (\theta - n)] \tag{9.78}$$

where $f^{[n+1]}(\xi)$ is the value of the $(n + 1)$st *derivative* of the function, assumed to be constant over the range $x_0 \leq \xi \leq x_n$. For Newton's backward-difference interpolation formula, the corresponding remainder is

$$R = \frac{h^{n+1}f^{[n+1]}(\xi)}{(n+1)!}[\theta(\theta + 1)(\theta + 2) \cdots (\theta + n)] \tag{9.79}$$

For the Lagrange formula,

$$R = \frac{(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n)}{(n+1)!} f^{[n+1]}(\xi) \tag{9.80}$$

When the spacing is equal, Eq. (9.80) reduces to Eq. (9.78), the Newton forward-difference interpolation formula. Remainder terms for these and the other interpolation formulas are listed in most texts on numerical analysis.

The *rounding off* of numbers is another source of error. Should this occur in the values of the given tabulated function values, there will result oscillations in the $(n + 1)$st differences of the difference table. This was demonstrated in Table 9.5. Rounding off in the individual terms of the interpolation formulas and in their sums will also contribute to error. This is especially true when the Lagrange formula is used.

9.5 *Inverse Interpolation*

Interpolation, the subject of Sec. 9.4, requires the determination of $f(x)$ for a particular x located between two tabulated values. Inverse interpolation is concerned with the problem of given $f(x)$, determine x.

Since for all but the most elementary situations the interval of the tabulated values of $y = f(x)$ is not constant, Lagrange's interpolation formula for unequal spacings can be directly applied.

$$x = \frac{(y - y_1)(y - y_2) \cdots (y - y_n)}{(y_0 - y_1)(y_0 - y_2) \cdots (y_0 - y_n)} x_0 + \frac{(y - y_0)(y - y_2) \cdots (y - y_n)}{(y_1 - y_0)(y_1 - y_2) \cdots (y_1 - y_n)} x_1 + \frac{(y - y_0)(y - y_1)(y - y_3) \cdots (y - y)_n}{(y_2 - y_0)(y_2 - y_1)(y_2 - y_3) \cdots (y_2 - y_n)} x_2 + \cdots \tag{9.81}$$

For those cases where the function values $f(x)$ are tabulated at equal increments of the argument x, an alternative, iterative procedure can be developed directly from the equations of Table 9.20. Consider, for illustration, the first of the interpolation formulas listed, Newton's forward-difference interpolation formula, which is here repeated:

$$f(x + \theta h) = f(x) + \theta\,\Delta f(x) + \frac{\theta(\theta - 1)}{2!}\Delta^2 f(x) + \cdots + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - n + 1)}{n!}\Delta^n f(x) \tag{9.82}$$

It will be assumed that $0 \le \theta \le +1$ and that the x position in Eq. (9.82) corresponds to the 0 location in the difference table. That is,

$$\begin{aligned} f(x) &= y_0 & \qquad \Delta^3 f(x) &= \Delta^3 y_0 \\ \Delta f(x) &= \Delta y_0 & \cdots &\cdots \\ \Delta^2 f(x) &= \Delta^2 y_0 \end{aligned} \tag{9.83}$$

In addition, the known function $f(x + \theta h)$ will be denoted as y_θ. Equation (9.82) can then be written as in Eq. (9.84) or (9.85).

$$y_\theta = y_0 + \theta\,\Delta y_0 + \frac{\theta(\theta - 1)}{2!}\Delta^2 y_0 + \cdots + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - n + 1)}{n!}\Delta^n y_0 \tag{9.84}$$

$$y_\theta - y_0 = \theta\,\Delta y_0 + \frac{\theta(\theta - 1)}{2!}\Delta^2 y_0 + \cdots + \frac{\theta(\theta - 1)(\theta - 2) \cdots (\theta - n + 1)}{n!}\Delta^n y_0 \tag{9.85}$$

A first approximation to θ can be obtained by terminating (or truncating) the series after the first difference Δy_0.

$$y_\theta - y_0 = {}_1\theta\,\Delta y_0 \qquad \text{or} \qquad {}_1\theta = \frac{y_\theta - y_0}{\Delta y_0} \tag{9.86}$$

By using this computed first approximation to θ, a second approximation can be obtained by including two terms of the series as indicated in Eq. (9.87).

$$y_\theta - y_0 = {}_2\theta\,\Delta y_0 + \frac{{}_1\theta({}_1\theta - 1)}{2!}\Delta^2 y_0 \tag{9.87}$$

or

$${}_2\theta = \frac{(y_\theta - y_0) - \dfrac{{}_1\theta({}_1\theta - 1)}{2}\Delta^2 y_0}{\Delta y_0} \tag{9.88}$$

The process is repeated according to Eqs. (9.89) until the change in θ between two successive iterations is equal to or less than the desired error.

$$\begin{aligned} y_\theta - y_0 &= {}_3\theta\,\Delta y_0 + \frac{{}_2\theta({}_2\theta - 1)}{2!}\Delta^2 y_0 + \frac{{}_2\theta({}_2\theta - 1)({}_2\theta - 2)}{3!}\Delta^3 y_0 \\ y_\theta - y_0 &= {}_4\theta\,\Delta y_0 + \frac{{}_3\theta({}_3\theta - 1)}{2!}\Delta^2 y_0 + \frac{{}_3\theta({}_3\theta - 1)({}_3\theta - 2)}{3!}\Delta^3 y_0 \\ &\quad + \frac{{}_3\theta({}_3\theta - 1)({}_3\theta - 2)({}_3\theta - 3)}{4!}\Delta^4 y_0 \end{aligned} \tag{9.89}$$

. . . .

As an example of inverse interpolation, consider the determination of x for $f(x) = 24{,}200$, given the difference Table 9.25. In terms of the nota-

Table 9.25

x	$y = f(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$
86	23,632			
		237		
87	23,869		3	
		240		−1
88	24,109		2	
		242		+1
89	24,351		3	
		245		−1
90	24,596		2	
		247		
91	24,843			

tion of Eqs. (9.85) through (9.89),

$$y_\theta = 24{,}200 \qquad y_0 = 24{,}109 \qquad \text{(corresponding to } x = 88\text{)}$$

$$\begin{aligned} \Delta y_0 &= 242 \\ \Delta^2 y_0 &= 3 \\ \Delta^3 y_0 &= -1 \end{aligned} \tag{9.90}$$

$$y_\theta - y_0 = 91$$

From Eq. (9.86),

$$91 = {}_1\theta(242) \qquad \text{or} \qquad {}_1\theta = 0.376 \tag{9.91}$$

From Eq. (9.87),

$$91 = {}_2\theta(242) + \frac{(0.376)(0.376 - 1.000)}{2}(3)$$

or

$${}_2\theta = 0.378 \tag{9.92}$$

Using the first of Eqs. (9.89),

$$91 = {}_3\theta(242) + \frac{(0.378)(0.378 - 1.000)}{2!}(3) + \frac{(0.378)(0.378 - 1.000)(0.378 - 2.000)}{3!}(-1)$$

which gives

$${}_3\theta = 0.378 \tag{9.93}$$

The value of x corresponding to $f(x) = 24{,}200$ is therefore

$$x_0 + \theta h = 88 + (0.378)(1) = 88.378 \tag{9.94}$$

9.6 Numerical Differentiation and Integration

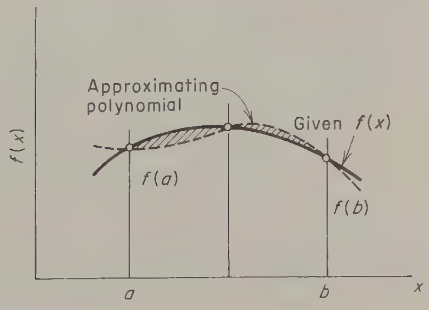
It has been shown in the preceding sections that continuous "well-ordered" functions can be conveniently treated even though they may be represented only in tabular form. With appropriate formulas, interpolation, extrapolation, and inverse interpolation can be carried out to any desired degree of accuracy. In reality, then, the formulas developed in Secs. 9.3 to 9.5 can be considered to actually represent the real $f(x)$, at least over a limited range of x.

Interpolation and extrapolation are not the only numerical processes required in engineering and scientific analysis. Differentiation and integration must also be performed. If the function in question is described only in tables, the process by which the value of the derivative of the function at a given location is determined is known as numerical differentiation. Similarly, by numerical integration, the value of the integral over a specified interval can be determined, even though the function is defined only at discrete points.

An obvious method for obtaining numerical approximations to the value of the derivative is to differentiate the interpolation formulas. As stated above, these represent to the desired accuracy the real $f(x)$. Numerical-integration formulas would be obtained by integrating these same interpolation formulas. If over the range of x in question the equations adequately represent the function values at the tabulated points, the developed numerical-differentiation and numerical-integration formulas should provide reasonable approximations to the desired functions.

In general, the process of numerical integration is stable and dependable. Such is not always the case for numerical differentiation. The situation is best illustrated by considering Fig. 9.3. Here it is assumed that a given function $f(x)$ is prescribed over an interval $a \leq x \leq b$. Moreover, values of the function are presumed to be specified at a, at b, and also at one intermediate point. In addition, there is shown on the figure an approximating polynomial, an interpolation equation, which passes through these same three given points. It is to be noted that the approximation underestimates the true value of the function in part of the range in question, and overestimates it in other parts. This is typical of the actual situation that is observed when approximating polynomials are used to represent given functions. Since $\int_a^b f(x)\,dx$ is equal to the area under $f(x)$ between a and b, and since the approximating polynomial both underestimates and overestimates, the errors in numerical integration tend to be compensating in nature. Truncation is therefore less of a problem. Numerical differentiation realizes no such benefit. It is obvious that the derivative, that is, the slope of the curve, of the approximating polyno-

Figure 9.3



DIFFERENCE BETWEEN
TABLE 9.18 & TABLES (9.26-28

mial may be markedly different from that of the given $f(x)$. This is especially true when $f(x)$ varies appreciably within the interval in question. Two methods can be used to improve accuracy in numerical differentiation. A higher-order approximating polynomial, one that requires a larger number of given pivotal points, can be introduced. The second method requires that the interval in x between the given pivotal points be reduced.

9.6.1 Difference Expressions

In Sec. 9.3 it was demonstrated that there exist relationships between the operator D (where $D = d/dx$) and the difference operators Δ and ∇. The specific relationships were listed in Tables 9.17 and 9.18. Similar expressions also could be developed relating D and δ. The important point is that the operator D is presumed to represent the derivative in a *continuous system*, whereas Δ, ∇, and δ correspond to values tabulated at a number of separate and distinct points.

It has also been demonstrated [see Eq. (9.41)] that the operator D can be expressed in terms of function values at known points. The exact relationship depends on the order of the approximation and the particular selection of interpolation formula.

A second-order approximating function, one where Δ^2, ∇^2, and δ^2 are constants and higher-order differences are presumed to be zero, can be obtained by passing through three consecutive points a single second-order polynomial of the type given in Eq. (9.60), here repeated as Eq. (9.95).

$$y = a_2x^2 + a_1x + a_0 \tag{9.95}$$

For symmetrically placed pivotal points, as indicated in Fig. 9.4,

$$\begin{aligned} a_0 &= y_0 \\ a_1 &= \frac{y_{+1} - y_{-1}}{2h} \\ a_2 &= \frac{y_{-1} - 2y_0 + y_{+1}}{2h^2} \end{aligned} \tag{9.96}$$

This gives, for the difference operator evaluated at $x = 0$,

$$D(y_0) = (2a_2x + a_1)_{x=0} = a_1$$

or

$$D(y_0) = \frac{y_{+1} - y_{-1}}{2h} \tag{9.97}$$

In like manner

$$D^2(y_0) = \frac{y_{-1} - 2y_0 + y_{+1}}{h^2} \tag{9.98}$$

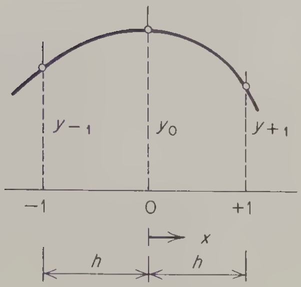


Figure 9.4

The corresponding second-order forward-difference expressions are

$$D(y_0) = \frac{-3y_0 + 4y_{+1} - y_{+2}}{2h} \tag{9.99}$$

$$D^2(y_0) = \frac{y_0 - 2y_{+1} + y_{+2}}{h^2} \tag{9.100}$$

POINT -
MATCH

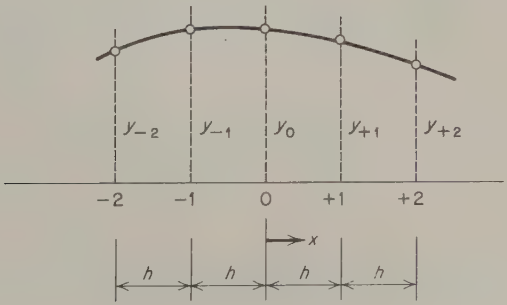
It should be noted that in both cases the second-difference expressions, Eqs. (9.98) and (9.100), are of exactly the same form. This is not the case for the first differences.

To illustrate the development of "higher-order" difference expressions, consider the fourth-order case with assumed central placement of the coordinate reference. (See Fig. 9.5.) The selected polynomial is

$$y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \tag{9.101}$$

Proceeding as before to evaluate a_4, a_3, a_2, a_1, and a_0 in terms of the assumed, known function values at locations −2, −1, 0, +1, and +2

Figure 9.5



yields the five linear equations (9.102):

$$\begin{bmatrix} y_{-2} \\ y_{-1} \\ y_0 \\ y_{+1} \\ y_{+2} \end{bmatrix} = \begin{bmatrix} +16h^4 & -8h^3 & +4h^2 & -2h & +1 \\ +h^4 & -h^3 & +h^2 & -h & +1 \\ 0 & 0 & 0 & 0 & +1 \\ +h^4 & +h^3 & +h^2 & +h & +1 \\ +16h^4 & +8h^3 & +4h^2 & +2h & +1 \end{bmatrix} \begin{bmatrix} a_4 \\ a_3 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} \tag{9.102}$$

The values of the constants become

POINT-
MATCH

$$\begin{aligned} a_4 &= \frac{+y_{-2} - 4y_{-1} + 6y_0 - 4y_{+1} + y_{+2}}{24h^4} \\ a_3 &= \frac{-y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2}}{12h^3} \\ a_2 &= \frac{-y_{-2} + 16y_{-1} - 30y_0 + 16y_{+1} - y_{+2}}{24h^2} \\ a_1 &= \frac{+y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2}}{12h} \\ a_0 &= y_0 \end{aligned} \tag{9.103}$$

With these, the derivatives evaluated at $x = 0$ can be readily defined.

$$\begin{aligned} y_0 &= a_0 \\ y_0' &= D(y_0) = a_1 \\ y_0'' &= D^2(y_0) = 2a_2 \\ y_0''' &= D^3(y_0) = 6a_3 \\ y_0'''' &= D^4(y_0) = 24a_4 \end{aligned} \tag{9.104}$$

Substituting the values of the constants into Eqs. (9.104) yields difference expressions for the first four derivatives.

$$\begin{aligned} D(y_0) &= \frac{+y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2}}{12h} \\ D^2(y_0) &= \frac{-y_{-2} + 16y_{-1} - 30y_0 + 16y_{+1} - y_{+2}}{12h^2} \\ D^3(y_0) &= \frac{-y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2}}{2h^3} \\ D^4(y_0) &= \frac{+y_{-2} - 4y_{-1} + 6y_0 - 4y_{+1} + y_{+2}}{h^4} \end{aligned} \tag{9.105}$$

It is evident that equations similar to these can equally well be developed in terms of either forward or backward differences.

As was true in the development of interpolation formulas, it is desirable to have an indication of the magnitude of the error associated with a given assumed polynomial approximation. Several methods for doing this are available; the one selected for illustration presupposes the expansion in a Taylor series of the assumed continuous function $f(x)$. That is,

$$f(x+h) = f(x) + \frac{h}{1!} D(y) + \frac{h^2}{2!} D^2(y) + \frac{h^3}{3!} D^3(y) + \frac{h^4}{4!} D^4(y) + \cdots \tag{9.106}$$

Expanding about point zero,

$$y_{+1} = y_0 + h\,D(y_0) + \frac{1}{2!} h^2 D^2(y_0) + \frac{1}{3!} h^3 D^3(y_0) + \frac{1}{4!} h^4 D^4(y_0) + \cdots \tag{9.107}$$

Similarly,

$$y_{-1} = y_0 - h\,D(y_0) + \frac{1}{2!} h^2 D^2(y_0) - \frac{1}{3!} h^3 D^3(y_0) + \frac{1}{4!} h^4 D^4(y_0) - \cdots \tag{9.108}$$

Adding corresponding members of Eqs. (9.107) and (9.108) yields

$$y_{+1} + y_{-1} = 2y_0 + \frac{2}{2!} h^2 D^2(y_0) + \frac{2}{4!} h^4 D^4(y_0) + \cdots \tag{9.109}$$

Solving for $D^2(y_0)$,

$$D^2(y_0) = \frac{y_{-1} - 2y_0 + y_{+1}}{h^2} - \frac{1}{12} h^2 D^4(y_0) + \cdots \tag{9.110}$$

This is frequently written as

$$D^2(y_0) = \frac{y_{-1} - 2y_0 + y_{+1}}{h^2} - \frac{1}{12} h^2 y_0'''' + \cdots \tag{9.111}$$

where the next nonvanishing term in the Taylor series, beyond that developed from the assumed second-order equation, is

$$-\tfrac{1}{12} h^2 y_0''''$$

If, instead of adding with Eqs. (9.107) and (9.108), subtracting is effected, the following results:

$$y_{+1} - y_{-1} = 2h\,D(y_0) + \frac{2}{3!} h^3 D^3(y_0) + \cdots \tag{9.112}$$

or

$$D(y_0) = \frac{y_{+1} - y_{-1}}{2h} - \frac{1}{6} h^2 y_0''' \cdots \tag{9.113}$$

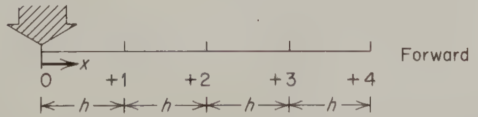
Difference expressions for forward, symmetrical, and backward placement of the pivotal locations are summarized in Tables 9.26 to 9.28.

FORWARD

Table 9.26†

Order of Eq.	Second order $y = a_2x^2 + a_1x + a_0$	Third order $y = a_3x^3 + a_2x^2 + a_1x + a_0$	Fourth order $y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$
$D(y_0)$	$\frac{1}{2h}(-3y_0 + 4y_{+1} - y_{+2})$ $+$ [redacted] $)$	$\frac{1}{6h}(-11y_0 + 18y_{+1} - 9y_{+2} + 2y_{+3})$ [redacted]	$\frac{1}{24h}(-50y_0 + 96y_{+1} - 72y_{+2} + 32y_{+3} - 6y_{+4})$ [redacted]
$D^2(y_0)$	$\frac{1}{h^2}(y_0 - 2y_{+1} + y_{+2})$ $-$ [redacted] $)$	$\frac{1}{h^2}(2y_0 - 5y_{+1} + 4y_{+2} - y_{+3})$ $+$ [redacted] $)$	$\frac{1}{12h^2}(35y_0 - 104y_{+1} + 114y_{+2} - 56y_{+3} + 11y_{+4})$ [redacted] $)$
$D^3(y_0)$		$\frac{1}{h^3}(-y_0 + 3y_{+1} - 3y_{+2} + y_{+3})$ $-$ [redacted]	$\frac{1}{4h^3}(-10y_0 + 36y_{+1} - 48y_{+2} + 28y_{+3} - 6y_{+4})$ $+$ [redacted]
$D^4(y_0)$			$\frac{1}{h^4}(y_0 - 4y_{+1} + 6y_{+2} - 4y_{+3} + y_{+4}) -$ [redacted]

Note: Next nonvanishing term in Taylor series is shown in parentheses.



---

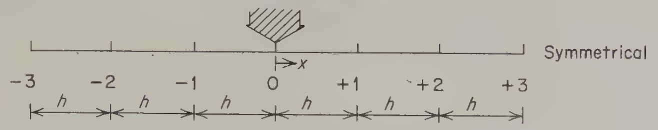
† Information contained in this table is from "Formulae for Numerical Differentiation," by W. G. Bickley, *Mathematical Gazette*, 1941, vol. 25, pp. 19–27. It is reproduced by permission of G. Bell and Sons, Ltd., London, England.

CENTRAL

Table 9.27† NEED 3 PTS NEED 5 PTS NEED 7 PTS

Order of Eq.	Second order $y = a_2x^2 + a_1x + a_0$	Fourth order $y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$	Sixth order $y = a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$
$D(y_0)$	$\frac{1}{2h}(-y_{-1} + y_{+1})$ ~~$+\left(-\frac{1}{6}h^2y_0'''\right)$~~	$\frac{1}{12h}(y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2})$ ~~$+\left(\frac{1}{30}h^4y_0^{V}\right)$~~	$\frac{1}{60h}(-y_{-3} + 9y_{-2} - 45y_{-1} + 45y_{+1} - 9y_{+2} + y_{+3})$ $+$ ~~$\left(-\frac{1}{140}h^6y_0^{VII}\right)$~~
$D^2(y_0)$	$\frac{1}{h^2}(y_{-1} - 2y_0 + y_{+1})$ $+$ ~~$\left(-\frac{1}{12}h^2y_0^{IV}\right)$~~	$\frac{1}{12h^2}(-y_{-2} + 16y_{-1} - 30y_0 + 16y_{+1} - y_{+2})$ ~~$+\left(\frac{1}{90}h^4y_0^{VI}\right)$~~	$\frac{1}{180h^2}(2y_{-3} - 27y_{-2} + 270y_{-1} - 490y_0 + 270y_{+1}$ $- 27y_{+2} + 2y_{+3}) +$ ~~$\left(-\frac{1}{560}h^6y_0^{VIII}\right)$~~
$D^3(y_0)$		$\frac{1}{2h^3}(-y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2})$ ~~$+\left(-\frac{1}{4}h^2y_0^{V}\right)$~~	$\frac{1}{8h^3}(y_{-3} - 8y_{-2} + 13y_{-1} - 13y_{+1} + 8y_{+2} - y_{+3})$ $+$ ~~$\left(\frac{7}{120}h^4y_0^{VII}\right)$~~
$D^4(y_0)$		$\frac{1}{h^4}(y_{-2} - 4y_{-1} + 6y_0 - 4y_{+1} + y_{+2})$ ~~$+\left(-\frac{1}{6}h^2y_0^{VI}\right)$~~	$\frac{1}{6h^4}(-y_{-3} + 12y_{-2} - 39y_{-1} + 56y_0 - 39y_{+1}$ $+ 12y_{+2} - y_{+3}) +$ ~~$\left(\frac{7}{240}h^4y_0^{VIII}\right)$~~

Note: Next nonvanishing term in Taylor series is shown in parentheses.



---

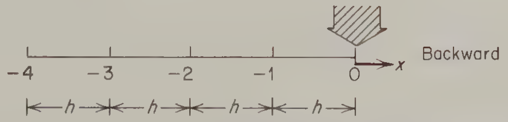
† Information contained in this table is from "Formulae for Numerical Differentiation," by W. G. Bickley, *Mathematical Gazette*, 1941, vol. 25, pp. 19–27. It is reproduced by permission of G. Bell and Sons, Ltd., London, England.

BACKWARD

Table 9.28†

Order of Eq.	Second order $y = a_2x^2 + a_1x + a_0$	Third order $y = a_3x^3 + a_2x^2 + a_1x + a_0$	Fourth order $y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$
$D(y_0)$	$\frac{1}{2h}(y_{-2} - 4y_{-1} + 3y_0)$ + [redacted]	$\frac{1}{6h}(-2y_{-3} + 9y_{-2} - 18y_{-1} + 11y_0)$ + [redacted]	$\frac{1}{24h}(6y_{-4} - 32y_{-3} + 72y_{-2} - 96y_{-1} + 50y_0)$ + [redacted]
$D^2(y_0)$	$\frac{1}{h^2}(y_{-2} - 2y_{-1} + y_0)$ + [redacted]	$\frac{1}{h^2}(-y_{-3} + 4y_{-2} - 5y_{-1} + 2y_0)$ + [redacted]	$\frac{1}{12h^2}(11y_{-4} - 56y_{-3} + 114y_{-2} - 104y_{-1} + 35y_0)$ + [redacted]
$D^3(y_0)$		$\frac{1}{h^3}(-y_{-3} + 3y_{-2} - 3y_{-1} + y_0)$ + [redacted]	$\frac{1}{4h^3}(6y_{-4} - 28y_{-3} + 48y_{-2} - 36y_{-1} + 10y_0)$ + [redacted]
$D^4(y_0)$			$\frac{1}{h^4}(y_{-4} - 4y_{-3} + 6y_{-2} - 4y_{-1} + y_0)$ + [redacted]

Note: Next nonvanishing term in Taylor series is shown in parentheses.



---

† Information contained in this table is from "Formulae for Numerical Differentiation," by W. G. Bickley, *Mathematical Gazette*, 1941, vol. 25, pp. 19–27. It is reproduced by permission of G. Bell and Sons, Ltd., London, England.

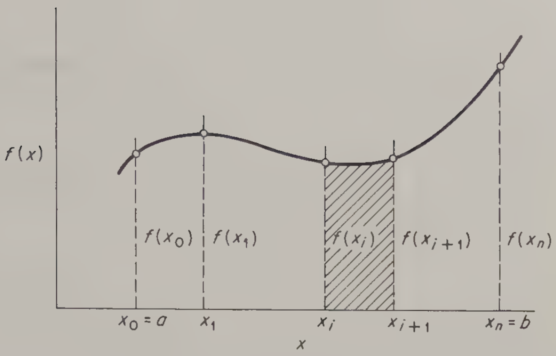


Figure 9.6

Second-, third-, and fourth-order polynomials are selected for both the forward and the backward cases. Second-, fourth-, and sixth-order expressions are presumed for the symmetrical cases. Each entry contains, in parentheses, the next nonvanishing term in the Taylor series.

9.6.2 Integration

Equations for numerical integration are normally developed by direct integration of interpolating polynomials. By assuming, for illustration, a general Lagrange interpolation formula, and using the notation shown in Fig. 9.6, the equation for $\int_a^b f(x)\,dx$ is

$$\int_a^b f(x)\,dx = \int_a^b \sum_{i=0}^{m} \prod_{\substack{j=0 \\ i\neq j}}^{m} \frac{x - x_j}{x_i - x_j} f(x_i)\,dx \tag{9.114}$$

Assuming the most elementary form of the polynomial—namely, a straight line passing through two points x_i, $f(x_i)$ and x_{i+1}, $f(x_{i+1})$—gives

$$\sum \Pi(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} f(x_i) + \frac{x - x_i}{x_{i+1} - x_i} f(x_{i+1}) \tag{9.115}$$

or
$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \left[\frac{f(x_i)}{x_i - x_{i+1}} \frac{(x - x_{i+1})^2}{2} + \frac{f(x_{i+1})}{x_{i+1} - x_i} \frac{(x - x_i)^2}{2} \right]_{x_i}^{x_{i+1}} \tag{9.116}$$

This gives

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \frac{f(x_i)}{-2h}(-1)(x_i - x_{i+1})^2 + \frac{f(x_{i+1})}{2h}(x_{i+1} - x_i)^2$$

or
$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \frac{h}{2}[f(x_i) + f(x_{i+1})] \tag{9.117}$$

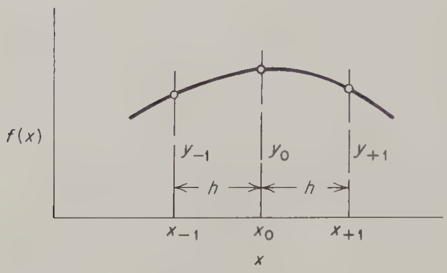


Figure 9.7

It is to be noted that the interval between x_i and x_{i+1} is presumed to equal h. This is the well-known *trapezoidal rule* for numerical integration.

More accuracy could be realized by assuming that the polynomial passed through three equally spaced points. For the notation given in Fig. 9.7, the Lagrange polynomial is given by

$$\sum \Pi(x) = \frac{(x - x_0)(x - x_1)}{(x_{-1} - x_0)(x_{-1} - x_1)} y_{-1} + \frac{(x - x_{-1})(x - x_1)}{(x_0 - x_{-1})(x_0 - x_1)} y_0 + \frac{(x - x_{-1})(x - x_0)}{(x_1 - x_{-1})(x_1 - x_0)} y_{+1} \tag{9.118}$$

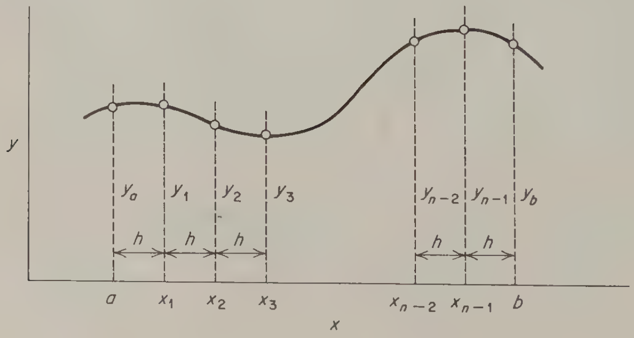
or

$$\sum \Pi(x) = \frac{x(x - h)}{2h^2} y_{-1} + \frac{(x + h)(x - h)}{-h^2} y_0 + \frac{(x + h)(x)}{2h^2} y_{+1} \tag{9.119}$$

or

$$\sum \Pi(x) = \frac{y_{-1} - 2y_0 + y_{+1}}{2h^2} x^2 + \frac{y_{+1} - y_{-1}}{2h} x + y_0 \tag{9.120}$$

Figure 9.8



This gives as a numerical-integration formula

$$\int_{-h}^{+h} f(x)\,dx = \int_{-h}^{+h} \sum \Pi(x)\,dx = \frac{h}{3}(y_{-1} + 4y_0 + y_{+1}) \qquad (9.121)$$

which is known as *Simpson's one-third rule* for numerical integration. By considering the entire range from $a \le x \le b$ (see Fig. 9.8), the integral can be expressed as

$$\int_a^b f(x)\,dx = \frac{h}{3}[(y_a + 4y_1 + y_2) + (y_2 + 4y_3 + y_4) + \cdots + (y_{n-2} + 4y_{n-1} + y_b)] \qquad (9.122)$$

$$\boxed{\text{or} \quad \int_a^b f(x)\,dx = \frac{h}{3}(y_a + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 2y_{n-2} + 4y_{n-1} + y_b) \qquad (9.123)}$$

It is to be noted that in this development the interval $a \le x \le b$ is subdivided into an "even" number of segments of length h.

The same procedure used in the development of the trapezoidal rule and Simpson's one-third rule, that is, starting from the Lagrange interpolation formula, can be used to develop higher-order equations.

As noted earlier in this section, any of the interpolation formulas can be used to develop integration equations. For completeness and to facilitate the development in a later chapter of methods of solution to differential equations, consider here also, then, the establishment of integration formulas from a direct evaluation of the finite-difference formulas listed in Table 9.20. (It should be noted that, in general, integration formulas of the type and kind considered in this volume are frequently referred to in other books and articles as "quadrature formulas.")

Newton's forward-difference interpolation formula (from Table 9.20) is defined by the following equation:

$$f(x + \theta h) = f(x) + \theta\,\Delta f(x) + \frac{\theta(\theta - 1)}{2!}\Delta^2 f(x) + \frac{\theta(\theta - 1)(\theta - 2)}{3!}\Delta^3 f(x) + \cdots \qquad (9.124)$$

If it is desired to evaluate the integral of the function between the limits x and $x + h$, that is, for the entire range between two consecutive pivotal points, integration is carried out with respect to θ, and θ is then set equal to 1.0

$$\int_x^{x+h} f(x)\,dx = \int_x^{x+\theta h} f(x)\,dx\,\Big|_{\theta = +1.0} = h\int_0^{+1} f(x + \theta)\,d\theta \qquad (9.125)$$

or

$$f(x)\,dx = h\int_0^{+1}\left[f(x) + \theta\,\Delta f(x) + \frac{\theta(\theta-1)}{2!}\,\Delta^2 f(x) + \frac{\theta(\theta-1)(\theta-2)}{3!}\,\Delta^3 f(x) + \cdots\right]d\theta$$

$$= h\left[\theta f(x) + \frac{\theta^2}{2}\,\Delta f(x) + \tfrac{1}{12}\,(2\theta^3 - 3\theta^2)\,\Delta^2 f(x) + \tfrac{1}{24}\,(\theta^4 - 12\theta^3 + 12\theta^2)\,\Delta^3 f(x) + \cdots\right]_0^{+1} \qquad (9.126)$$

Substituting in the limits on the integration gives the desired formula.

$$\int_x^{x+h} f(x)\,dx = h[f(x) + \tfrac{1}{2}\,\Delta f(x) + \tfrac{1}{12}\,\Delta^2 f(x) + \tfrac{1}{24}\,\Delta^3 f(x) - \tfrac{19}{720}\,\Delta^4 f(x) + \tfrac{3}{160}\,\Delta^5 f(x) + \cdots] \qquad (9.127)$$

Newton's backward-difference interpolation formula (from Table 9.20) is given as Eq. (9.128).

$$f(x + \theta h) = f(x) + \theta\,\nabla f(x) + \frac{\theta(\theta+1)}{2!}\,\nabla^2 f(x) + \frac{\theta(\theta+1)(\theta+2)}{3!}\,\nabla^3 f(x) + \cdots \qquad (9.128)$$

The corresponding integration formula is

$$\int_x^{x+h} f(x)\,dx = h[f(x) + \tfrac{1}{2}\,\nabla f(x) + \tfrac{5}{12}\,\nabla^2 f(x) + \tfrac{3}{8}\,\nabla^3 f(x) + \tfrac{251}{720}\,\nabla^4 f(x) + \tfrac{95}{288}\,\nabla^5 f(x) + \cdots] \qquad (9.129)$$

For the range of integration from x_0 to x_1, Bessel's and Everett's interpolation formulas give the same central-difference integration formulas.

$$\int_{x_0}^{x_1} f(x)\,dx = h\left(\frac{f_0 + f_1}{2} - \frac{1}{12}\,\frac{\delta^2 f_0 + \delta^2 f_1}{2} + \frac{11}{720}\,\frac{\delta^4 f_0 + \delta^4 f_1}{2} + \cdots\right) \qquad (9.130)$$

It is to be noted that the first term of this series corresponds identically to the trapezoidal rule developed earlier in this section by using the Lagrange interpolation formula.

It is frequently desirable to express integration formulas not in terms of differences but in terms of the function values at the fixed pivotal locations. For example, if only the first two terms of Eq. (9.130) are considered, then by noting that

$$\delta^2 f_0 = f_{-1} - 2f_0 + f_{+1} \qquad \text{and} \qquad \delta^2 f_1 = f_0 - 2f_{+1} + f_{+2}$$

the following integration formula can be obtained:

$$\int_{x_0}^{x_1} f(x)\,dx = h\left(\frac{-f_{-1} + 13f_0 + 13f_{+1} - f_{+2}}{24}\right) \tag{9.131}$$

Integration formulas expressed in terms of function values at pivotal points, rather than as differences of these, are said to be in *Lagrangian form.*

Integration need not be restricted to one interval. Similar equations can be developed for many "ranges of integration." Tables 9.29 and 9.30 summarize many of the more often referred to equations of this type. There also are given the orders of the corresponding polynomials and the next terms in the series associated with each. It is to be emphasized that these equations assume a constant interval of h between each of the pivotal locations.

9.7 Least-squares Approximations

STUDY

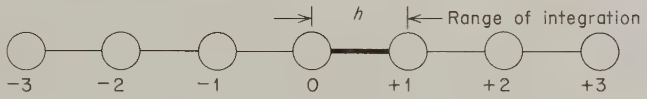
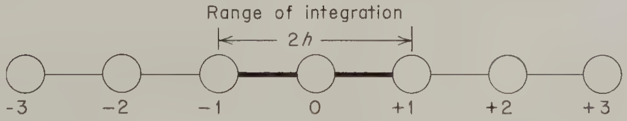
In the preceding sections of this chapter it is assumed as a necessary condition that the approximating polynomials must take on the values of the given function at the specified pivotal points. A natural consequence of this presumption is that the nth-order polynomial approximating equation requires exactly $n + 1$ pivotal points for evaluation.

In dealing with experimental data or in developing other types or orders of approximating equations, it is often desirable to include in the consideration a larger number of points than that which would be dictated by direct application of the hereinbefore developed polynomial approximations. For such cases it is desirable to specify at the outset, not the necessity that the polynomial pass through a given point or set of points, but the necessity that the sum of the errors at these points (or the sum of the squares of the errors between the assumed and the actual values) be minimized. In such a case a "best-fitting" relationship is realized.

The method of least squares is based on the assumption that the sum of the squares of the deviations between the assumed functions and the given data, that is, the sum of the squares of the errors, must be a minimum. If, as shown in Fig. 9.9, it is assumed that the ordinate of the data point is designated as y_i and the value of the approximating function at that same value of x_i is $f(x_i)$, the error at location x_i can be written as

$$\epsilon_i = y_i - f(x_i) \tag{9.132}$$

Table 9.29 Symmetrical Integration Formulas (in Lagrangian Form)

Equation number	Order of assumed polynomial	Equation	Next term in series
		 h Range of integration −3 −2 −1 0 +1 +2 +3	
1	Linear	$\int f(x)\,dx = h\dfrac{f_0 + f_1}{2}$	$-\dfrac{1}{12}h^3 f''(\xi)$
2	Third	$\int f(x)\,dx = h\dfrac{-f_{-1} + 13f_0 + 13f_{+1} - f_{+2}}{24}$	$+\dfrac{11}{720}h^5 f''''(\xi)$
3	Fifth	$\int f(x)\,dx = h\dfrac{11f_{-2} - 93f_{-1} + 802f_0 + 802f_{+1} - 93f_{+2} + 11f_{+3}}{1{,}440}$	$-\dfrac{191}{60{,}480}h^7 f^{\mathrm{VI}}(\xi)$
		 Range of integration 2h -3 −2 −1 0 +1 +2 +3	
4	Second	$\int f(x)\,dx = 2h\dfrac{f_{-1} + 4f_0 + f_{+1}}{6}$	$-\dfrac{1}{90}h^5 f''''(\xi)$
5	Fourth	$\int f(x)\,dx = 2h\dfrac{-f_{-2} + 34f_{-1} + 114f_0 + 34f_{+1} - f_{+2}}{180}$	$+\dfrac{1}{756}h^7 f^{\mathrm{VI}}(\xi)$
6	Sixth	$\int f(x)\,dx = 2h\dfrac{+5f_{-3} - 72f_{-2} + 1{,}503f_{-1} + 4{,}688f_0 + 1{,}503f_{+1} - 72f_{+2} + 5f_{+3}}{7{,}560}$	$-\dfrac{23}{113{,}400}h^9 f^{\mathrm{VIII}}(\xi)$

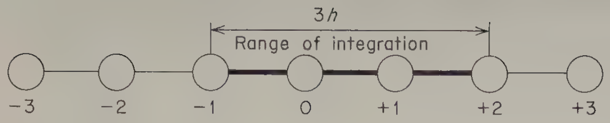
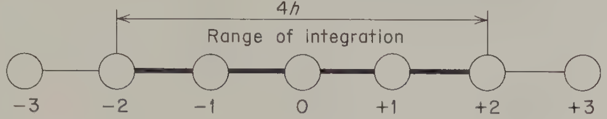
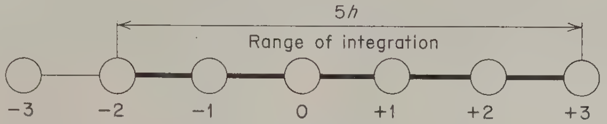
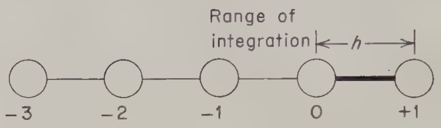
		 3h Range of integration −3 −2 −1 0 +1 +2 +3	
7	Linear	$\int f(x)\,dx = 3h\dfrac{f_0 + f_{+1}}{2}$	$+\dfrac{3}{4}h^3f''(\xi)$
8	Third	$\int f(x)\,dx = 3h\dfrac{f_{-1} + 3f_0 + 3f_{+1} + f_{+2}}{8}$	$-\dfrac{3}{80}h^5f''''(\xi)$
9	Fifth	$\int f(x)\,dx = 3h\dfrac{-f_{-2} + 23f_{-1} + 58f_0 + 58f_{+1} + 23f_{+2} - f_{+3}}{160}$	$+\dfrac{13}{2{,}240}h^7f^{\mathrm{VI}}(\xi)$
		 4h Range of integration −3 −2 −1 0 +1 +2 +3	
10	Second	$\int f(x)\,dx = 4h\dfrac{2f_{-1} - f_0 + 2f_{+1}}{3}$	$+\dfrac{14}{45}h^5f''''(\xi)$
11	Fourth	$\int f(x)\,dx = 4h\dfrac{7f_{-2} + 32f_{-1} + 12f_0 + 32f_{+1} + 7f_{+2}}{90}$	$-\dfrac{8}{945}h^7f^{\mathrm{VI}}(\xi)$
12	Sixth	$\int f(x)\,dx = 4h\dfrac{-4f_{-3} + 171f_{-2} + 612f_{-1} + 332f_0 + 612f_{+1} + 171f_{+2} - 4f_{+3}}{1{,}890}$	$+\dfrac{13}{14{,}175}h^9f^{\mathrm{VIII}}(\xi)$
		 5h Range of integration −3 −2 −1 0 +1 +2 +3	
13	Third	$\int f(x)\,dx = 5h\dfrac{11f_{-1} + f_0 + f_{+1} + 11f_{+2}}{24}$	$+\dfrac{95}{144}h^5f''''(\xi)$
14	Fifth	$\int f(x)\,dx = 5h\dfrac{19f_{-2} + 75f_{-1} + 50f_0 + 50f_{+1} + 75f_{+2} + 19f_{+3}}{288}$	$-\dfrac{275}{12{,}096}h^7f^{\mathrm{VI}}(\xi)$

Table 9.30 Nonsymmetrical (Last-interval) Integration Formulas (in Lagrangian Form)

Equation number	Order of assumed polynomial	Equation	Next term in series
			
1	Second	$\int f(x)\,dx = h\frac{-f_{-1} + 8f_0 + 5f_{+1}}{12}$	$+\frac{1}{24}h^5 f''''(\xi)$
2	Third	$\int f(x)\,dx = h\frac{f_{-2} - 5f_{-1} + 19f_0 + 9f_{+1}}{24}$	$-\frac{19}{720}h^6 f^{\mathrm{V}}(\xi)$
3	Fourth	$\int f(x)\,dx = h\frac{-19f_{-3} + 106f_{-2} - 264f_{-1} + 646f_0 + 251f_{+1}}{720}$	$+\frac{27}{1{,}440}h^7 f^{\mathrm{VI}}(\xi)$

The method of least squares requires that

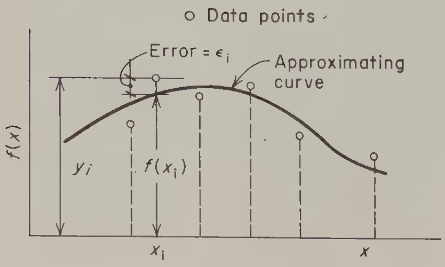
$$\sum_{i=1}^{m} [y_i - f(x_i)]^2 = \text{minimum} \tag{9.133}$$

where m is the number of data points.

For illustration, assume that the second-order polynomial, Eq. (9.134), is selected as the approximating function.

$$f(x) = a_2x^2 + a_1x + a_0 \tag{9.134}$$

Figure 9.9



The error at each of the "data points" is therefore

$$\text{Error} = (a_2x_i^2 + a_1x_i + a_0 - y_i) \qquad (i = 1, 2, \ldots, m)$$

The sum of the errors squared is

$$S = \Sigma(\text{error})^2 = \sum_{i=1}^{m} (a_2x_i^2 + a_1x_i + a_0 - y_i)^2 \tag{9.135}$$

For this to be a minimum it is necessary (but not sufficient) that

$$\frac{\partial S}{\partial a_2} = 0 \qquad \frac{\partial S}{\partial a_1} = 0 \qquad \frac{\partial S}{\partial a_0} = 0 \tag{9.136}$$

or

$$\begin{aligned} &\sum_{i=1}^{m} 2x_i^2(a_2x_i^2 + a_1x_i + a_0 - y_i) = 0 \\ &\sum_{i=1}^{m} 2x_i(a_2x_i^2 + a_1x_i + a_0 - y_i) = 0 \\ &\sum_{i=1}^{m} 2(a_2x_i^2 + a_1x_i + a_0 - y_i) = 0 \end{aligned} \tag{9.137}$$

Equations (9.137) represent a set in the unknowns a_0, a_1, and a_2. Recasting these in matrix form gives

$$\begin{bmatrix} \sum_{i=1}^{m} x_i^4 & \sum_{i=1}^{m} x_i^3 & \sum_{i=1}^{m} x_i^2 \\ \sum_{i=1}^{m} x_i^3 & \sum_{i=1}^{m} x_i^2 & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i^2 & \sum_{i=1}^{m} x_i & m \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_i^2y_i \\ \sum_{i=1}^{m} x_iy_i \\ \sum_{i=1}^{m} y_i \end{bmatrix} \tag{9.138}$$

As a numerical example, consider the determination by the method of least squares of the best-fitting second-order polynomial to the data given in Table 9.31. The calculations are shown in tabular form.

Table 9.31

x	y
0	0
1	0.500
2	0.866
3	1.000

x	x^2	x^3	x^4	y	xy	x^2y
0	0	0	0	0	0	0
1	1	1	1	0.500	0.500	0.500
2	4	8	16	0.866	1.732	3.464
3	9	27	81	1.000	3.000	9.000
$\sum_{i=1}^{m} = 6$	14	36	98	2.366	5.232	12.964

The corresponding matrix is

$$\begin{bmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 12.964 \\ 5.232 \\ 2.366 \end{bmatrix} \tag{9.139}$$

Solution gives the values of the coefficients of the polynomial as

$$a_2 = -0.0915 \qquad a_1 = 0.6111 \qquad a_0 = -0.0049 \tag{9.140}$$

The approximating polynomial is, therefore,

$$f(x) = -0.0915x^2 + 0.6111x - 0.0049 \tag{9.141}$$

Prob. 9.1: Determine any errors that may be present in the following table:

x	$f(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$
1.00	0.3413	0118	-6	0
1.05	0.3531	0112	-6	0
1.10	0.3643	0106	-1	5
1.15	0.3749	0105	-60	-59
1.20	0.3899	0045	43	103
1.25	0.3944	0088	-5	-48
1.30	0.4032	0083	-6	-1
1.35	0.4115	0077	-4	+2
1.40	0.4192	0073	-6	-2
1.45	0.4265	0067		0
1.50	0.4332			

Prob. 9.2: By interpolation, determine the value of $J_0(x)$ for $x = 9.736$.

$J_0(9.736) = -0.2259$

x	$J_0(x)$
9.5	−0.1939
9.6	−0.2090
9.7	−0.2218
9.8	−0.2323
9.9	−0.2403
10.0	−0.2459

Prob. 9.3: For the table given in Prob. 9.2, determine x corresponding to $J_0(x) = -0.2300$.

Prob. 9.4: The following tabulated data are given:

x	$f(x)$
−2	+8.000
−1	0.000
0	−2.050
+1	+1.894

Determine by each of the following methods the value of $f(x)$ at $x = +0.500$:

(*a*) Newton's backward-difference interpolation formula
(*b*) Gauss' backward-interpolation formula
(*c*) Everett's interpolation formula
(*d*) Method of least squares

Prob. 9.5: Construct divided-difference tables for each of the functions A_n and B_n.

n	A_n	B_n
2	0.416667	0.083333
3	0.375000	0.041667
5	0.329861	0.018750
7	0.304225	0.011367
10	0.280190	0.006787

Prob. 9.6: For the data contained in Prob. 9.5, develop Lagrange interpolation formulas for A_n and B_n.

Prob. 9.7: For the data contained in Prob. 9.5, develop interpolation polynomials by using the method of least squares.

Prob. 9.8: Determine the best-fitting polynomial of the form

$$y = a_2x^2 + a_1x + a_0$$

for the data

x	0	0.2	0.4	0.6	0.8	1.0
y	0.120	0.090	0.097	0.061	0.065	0.043

Prob. 9.9: Using the approximating polynomial

$$f(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

and assuming a symmetrical placement of the pivotal points as shown in Fig. 9.5, determine the numerical-integration formula corresponding to

$$\int_{-2h}^{+2h} f(x)\,dx$$

Prob. 9.10: Assuming forward differences and a sixth-order system, show that

$$\int_0^{6h} f(x)\,dx = \tfrac{3}{10}h(f_0 + 5f_1 + f_2 + 6f_3 + f_4 + 5f_5 + f_6)$$

Prob. 9.11: Show that

$$\int_0^{4h} f(x)\,dx = \tfrac{2}{45}h(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4)$$

Prob. 9.12: Numerically integrate e^{-x} between $0 \leq x \leq 1$, in steps of 0.1, by each of the following:

(*a*) Trapezoidal rule (*b*) Simpson's one-third rule

Prob. 9.13: To ascertain the error associated with the use of the trapezoidal rule for numerical integration, determine in the same manner as was done for numerical differentiation the next term in a Taylor series expansion solution.

Prob. 9.14: For the data contained in Prob. 9.2, determine the value of the derivative of the function $J_0(x)$ at $x = 9.736$.

Prob. 9.15: For the data contained in Prob. 9.5, determine the value of the second derivative of A_n with respect to n at $n = 5$.

Prob. 9.16: Given the function $y = x^3$, and the corresponding table

x	y
-2	-8.0
-1	-1.0
0	0
$+1$	$+1.0$
$+2$	$+8.0$

determine by interpolation the values of x^3 at

(*a*) -1.60 (*b*) $+0.05$ (*c*) $+1.50$

Prob. 9.17: By forward differences, determine the value of the second derivative of sin x at

(*a*) $x = \pi/2$ (*b*) $x = 0$

Assume that values of sin x are available at $x = 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$, and π. Compare the solution with exact values.

9.8 Computer Solutions

As pointed out in the introduction, a variety of types and methods of numerical analysis are considered in this chapter. Each could have associated with it a detailed flow diagram. It is to be noted, however, that many of the methods are computationally the same. Therefore, there is discussed in this section only the development of computer programs for

1. Construction of a forward-difference table
2. Development of the Lagrange interpolation formula
3. Numerical integration using Simpson's one-third rule

9.8.1 Construction of a Forward-difference Table

It is presumed that the table starts at some value x_k, which has a corresponding y_k, and proceeds to x_n, y_n. (*Note:* The table does not start necessarily at $k = 0$.) To facilitate computation there will be introduced a double subscript, for example. y_{ij}. For the case chosen i denotes the reference location in the table, and j denotes the order of the difference under examination.

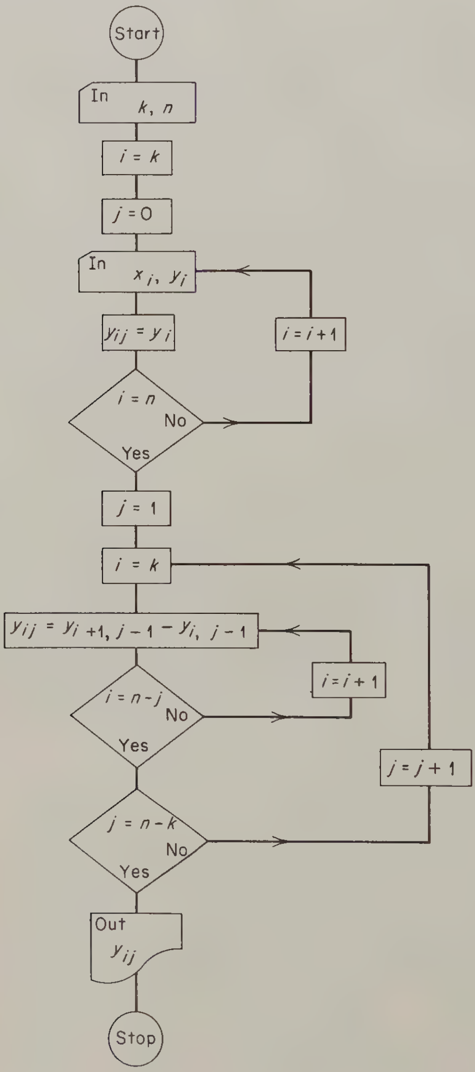
$$k \leq i \leq n$$

$$0 \leq j \leq \text{order of difference table desired}$$

It is to be noted that a 0 value for j denotes the function itself.

The flow diagram is given in Fig. 9.10.

Figure 9.10



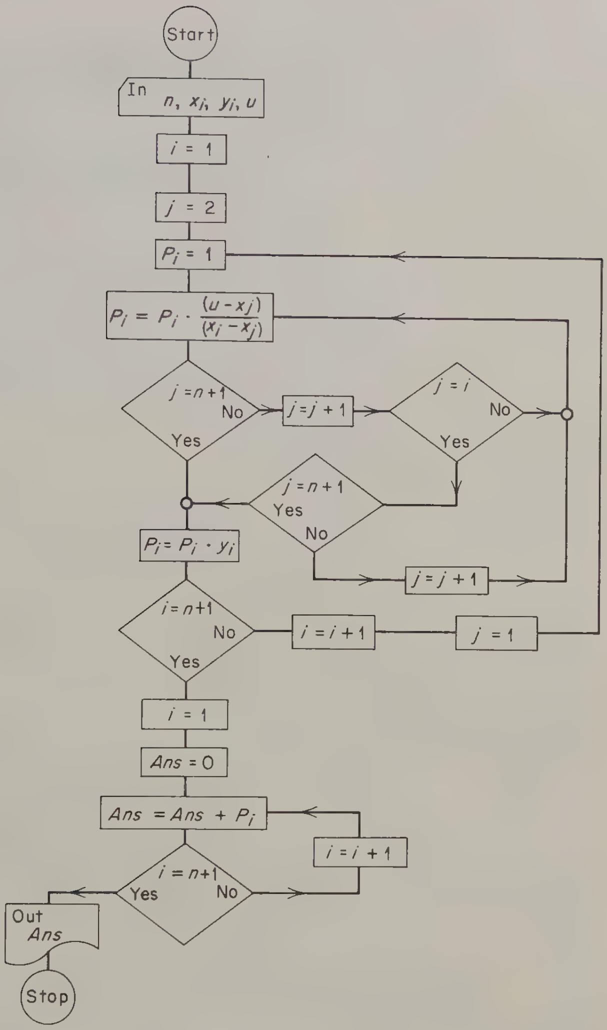


Figure 9.11

9.8.2 Development of the Lagrange Interpolation Formula

This method is well suited to computer solution since it is basically a repetitive process. The only problem encountered is that care must be exercised to ensure that any given stage i must not equal j. The value of x for which the interpolated value is desired is denoted by the letter u. The interpolated value is Ans. The flow diagram is given in Fig. 9.11.

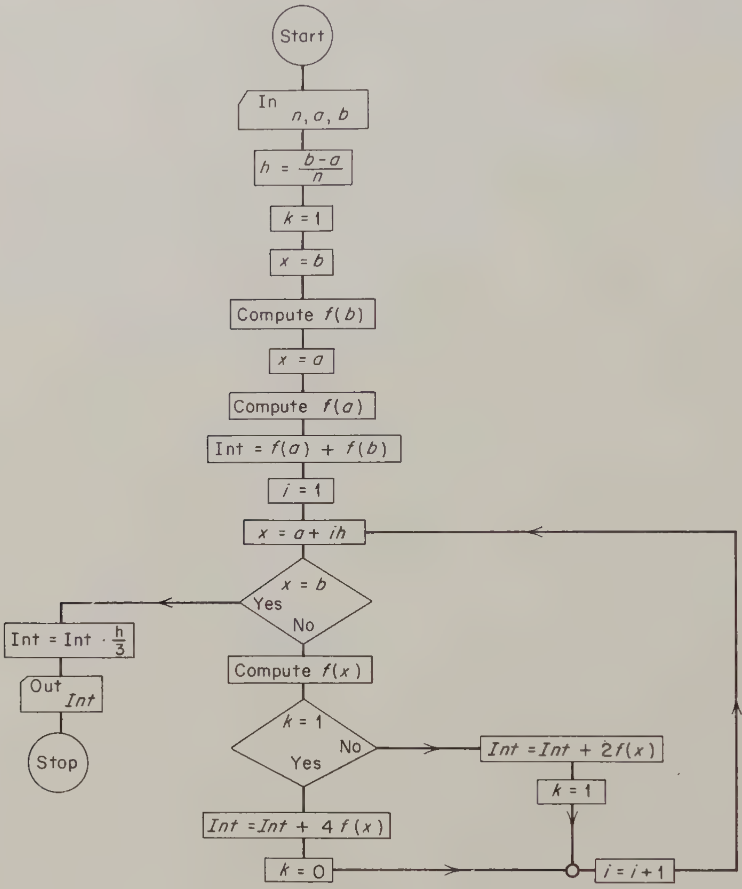


Figure 9.12

9.8.3 Numerical Integration Using Simpson's One-third Rule

The interval in question is $a \leq x \leq b$. It is assumed that the interval is divided into an even number of divisions n. The spacing of these divisions is denoted by h. In the case illustrated (Fig. 9.12), it is assumed that the function values can be computed, and need not be tabulated.

10 numerical solution of ordinary differential equations

10.1 Introduction

An equation containing one or more derivatives is called a *differential equation.* If these are ordinary derivatives, the equation is said to be an *ordinary differential equation;* if these are partial derivatives, the equation is termed a *partial differential equation.*

A differential equation which is of the first degree in the dependent variable and all its derivatives is termed a *linear equation.* Correspondingly, an equation in which the dependent variable or any of its derivatives is of the second or higher degree (or, more complexly, one in which the dependent variable is an argument of a transcendental function or the like) is termed a *nonlinear equation.* The order of an equation is the order of the highest derivative contained therein.

One means of representing a general ordinary linear differential equation in the dependent variable y and the independent variable x is shown in Eq. (10.1).

$$\sum_{i=0}^{n} A_i(x) \frac{d^i y}{dx^i} + A_{n+1}(x) = 0 \tag{10.1}$$

In this case, the coefficients $A_{n+1}(x)$ and $A_i(x)$ are shown as functions of x. The equation is therefore a linear ordinary differential equation with variable coefficients. Any or all the A terms may be constants.

An example of a second-order nonlinear variable-coefficient ordinary differential equation is given in Eq. (10.2).

$$\frac{d^2y}{dx^2} + e^x \left(\frac{dy}{dx}\right)^3 + y(\sin y) = x \tag{10.2}$$

Equation (10.3) is a second-order linear partial differential equation in the independent variables x and y and in the dependent variable z:

$$A_1(x,y)\frac{\partial^2 z}{\partial x^2} + A_2(x,y)\frac{\partial^2 z}{\partial x\,\partial y} + A_3(x,y)\frac{\partial^2 z}{\partial y^2} + A_4(x,y)\frac{\partial z}{\partial x} + A_5(x,y)\frac{\partial z}{\partial y} + A_6(x,y) + A_7(x,y)z = 0 \tag{10.3}$$

Here again, the coefficients have been shown as general functions of the independent variables, but any or all of these can be constants. When $A_6(x,y) = 0$, the equation is termed a *homogeneous* one.

In most cases, the analysis of problems in engineering and applied science requires that the dependent variable and/or its derivatives have certain specified magnitudes or interrelationships at prescribed values of the independent variable(s). Should all these constraints be imposed at a zero value of the independent variable(s), or should it be possible to transform the given equation to one in which this is true, the problem is said to be an *initial-value problem*. Should any of the conditions be imposed at other than zero, it is the *boundary-value* type. It is to be noted that problems of the mixed type can, and do, exist. One example is a transient heat-flow problem. In this case, spatial constraints are imposed at geometrical boundaries, and these may or may not be functions of time. In addition, the *initial state*, or initial values, of the problem would be normally specified at the instant of zero time. Since the exact general solution of an nth-order differential equation has n constants of integration, to ensure a unique solution there must be n independent boundary or initial conditions associated with each problem.

Relatively few "practical" differential equations can be solved in closed form. More often, solutions must be obtained by numerical procedures. The remainder of this and the following chapter will deal with a few of these methods.

Most numerical solutions of differential equations proceed from one of two approaches: series expansion or determining the undefined coefficients that result in the least error of an assumed approximate solution. Concerning the first of these, two basically different procedures can be followed to obtain the series in question. First, a general Taylor series expansion can be written, including the specified initial conditions and the values of the dependent derivatives of the function at that point. This is discussed in Sec. 10.2. The alternative procedure for obtaining a series representation is to iterate successively the corresponding integral equation, as described in Sec. 10.3. With regard to the second general approach, the determination of undefined coefficients based on error considerations, several different definitions of the significance of error can be considered, and four of these are discussed in Sec. 10.4.

10.2 Taylor Series Expansion: Initial-value Problems

The Taylor series expansion of a function in the vicinity of x_0 is

$$y(x) = y(x_0) + (x - x_0)y'(x_0) + \frac{(x - x_0)^2}{2!} y''(x_0) + \cdots \tag{10.4}$$

By denoting the interval from the original point of expansion x_0 to the general point x as h, Eq. (10.4) can be written in the summation form

$$y(x_0 + h) = \sum_{n=0} \frac{(h)^n}{n!} \left(\frac{d^n y}{dx^n}\right)_{x=x_0} \tag{10.5}$$

where $0! = 1$. Given then the values of the function and its derivatives at a particular point, say x_0, the solution for other values of x can be determined by suitably altering h.

10.2.1 First-order Differential Equations

To illustrate the general method of solution using the Taylor series expansion, consider the first-order differential equation (10.6), with the boundary condition as specified in Eq. (10.7).

$$\frac{dy}{dt} - 0.4y^3 = 1.0 \tag{10.6}$$

$$y(1.0) = 1.0 \tag{10.7}$$

Since the given differential equation is of the first order, only one independent boundary or initial condition can be specified [Eq. (10.7)]. Additional dependent higher-order derivatives can, however, be established, as is demonstrated.

By substituting Eq. (10.7) in (10.6), the value of the first derivative of y with respect to t at $t = 1.0$ can be determined. It is to be noted that this is not an independent condition, but is a consequence of the given original equation and the one given boundary condition.

$$\left(\frac{dy}{dt}\right)_{t=1.0} = \dot{y}(1.0) = 1.0 + 0.4 = 1.4 \tag{10.8}$$

By successively differentiating Eq. (10.6), the required values of the higher-order derivatives at $t = 1.0$ can also be established, one at a time.

$$\begin{aligned}
\left.\frac{dy}{dt}\right|_{t=1} &= 1.0 + 0.4y^3\Big|_{t=1} = 1.0 + 0.4(1.0)^3 = +1.4 \\
\left.\frac{d^2y}{dt^2}\right|_{t=1} &= 1.2y^2\frac{dy}{dt} = 1.2(1.0)^2(1.4)\Big|_{t=1} = +1.68 \\
\left.\frac{d^3y}{dt^3}\right|_{t=1} &= 2.4y\left(\frac{dy}{dt}\right)^2 + 1.2y^2\left.\frac{d^2y}{dt^2}\right|_{t=1} = +6.720 \\
\left.\frac{d^4y}{dt^4}\right|_{t=1} &= 2.4\left(\frac{dy}{dt}\right)^3 + 7.2y\frac{dy}{dt}\frac{d^2y}{dt^2} + 1.2y^2\left.\frac{d^3y}{dt^3}\right|_{t=1} = +31.5840 \\
&\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{10.9}$$

To the number of derivatives calculated above, the Taylor series describing this problem is given as

$$y = 1.0 + \frac{h}{1}(1.4) + \frac{h^2}{2!}(1.68) + \frac{h^3}{3!}(6.720) + \frac{h^4}{4!}(31.5840) + \cdots \tag{10.10}$$

or

$$y = 1.0 + 1.4h + 0.84h^2 + 1.120h^3 + 1.3160h^4 + \cdots \tag{10.11}$$

It is to be understood that the correct solution to the problem corresponds to the infinite series.

Several observations concerning Eq. (10.11) are readily apparent. Should h be less than 1.0, the series in all likelihood converges. Should a certain accuracy in y be desired, the smaller the interval of h that is being considered, the fewer the number of terms that are required in the terminated series. This is shown in Fig. 10.1, where the difference in using a second-order polynomial as opposed to a fourth-order equation at $h = 0.2$ is less than 1 percent. At $h = 0.1$, possibly just a linear approximation would suffice.

Another observation to be made with regard to the use of the Taylor series expansion is that if the differential equation is of second order, and if boundary conditions are specified at two separate values of the independent variable, the matching of this second given condition requires considerable additional work.

Suppose now that it is desired to establish the function $y(t)$ as defined by Eq. (10.6) in the range $1.0 \leq t \leq 2.0$. If it is presumed that a second-order polynomial is sufficient to give the accuracy needed for an interval

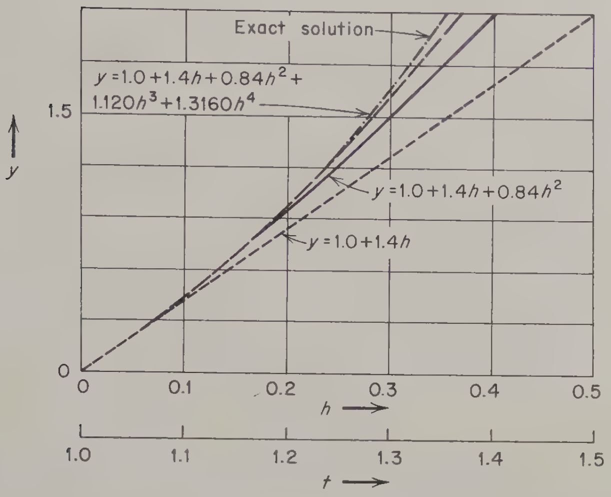


Figure 10.1

of $h = 0.2$, the problem can be considered not as just one large expansion over the total range in question, using $t = 1.0$ as the point of expansion, but rather as five successive smaller expansions: $1.0 \leq t \leq 1.2$, $1.2 \leq t \leq 1.4$, $1.4 \leq t \leq 1.6$, $1.6 \leq t \leq 1.8$, and $1.8 \leq t \leq 2.0$. In each range the necessary equations are

$$y(t) = y(t_0) + (t - t_0)y'(t_0) + \frac{(t - t_0)^2}{2} y''(t_0) \tag{10.12}$$

where t_0 corresponds to the starting value in that particular range in question. Similarly,

$$y'(t_0) = 1.0 + 0.4\,[y(t_0)]^3 \tag{10.13}$$

and

$$y''(t_0) = 1.2\,\{[y(t_0)]^2 y'(t_0)\} \tag{10.14}$$

It is to be noted that all calculations in one subregion are needed prior to starting the next set. Table 10.1 summarizes the results of these calculations for the problem in question. The corresponding curve is plotted in Fig. 10.2. There is also shown on the figure the exact solution of the differential equation.

Table 10.1 and Fig. 10.2 suggest a still further modification of the general Taylor series approach as thus far discussed. At the starting point of the first range (at $t = 1.0$), no information about the function is given for $t < 1.0$, and it is necessary to determine values of the higher-order interrelated boundary conditions in order to be able to effect the desired accuracy. The situation at $t = 1.2$ is not the same. Would it not be just as desirable, if not more so, to have the second range formulated so

Table 10.1

Range	t_0	$y(t_0)$	$y'(t_0)$	$y''(t_0)$	Equation of terminated Taylor series	$y(t_0 + 0.2)$
1 $1.0 < t < 1.2$ $(0 < h < 0.2)$	+1.0	+1.0	+1.4	+1.68	$y = 1.0 + 1.4h + 0.84h^2$	+1.314
2 $1.2 < t < 1.4$ $(0 < h < 0.2)$	+1.2	+1.314	+1.908	+3.952	$y = 1.314 + 1.908h + 1.976h^2$	+1.775
3 $1.4 < t < 1.6$ $(0 < h < 0.2)$	+1.4	+1.775	+3.237	+12.238	$y = 1.775 + 3.237h + 6.119h^2$	+2.667
4 $1.6 < t < 1.8$ $(0 < h < 0.2)$	+1.6	+2.667	+8.588	+64.767	$y = 2.667 + 8.588h + 32.384h^2$	+5.680
5 $1.8 < t < 2.0$ $(0 < h < 0.2)$	+1.8	+5.680	+74.300	+2,876.52	$y = 5.680 + 74.300h + 1{,}438.26h^2$	+78.070

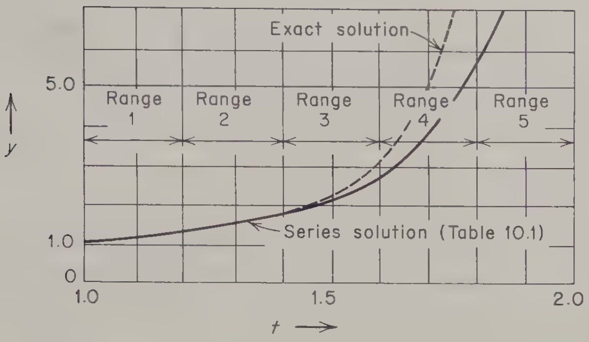
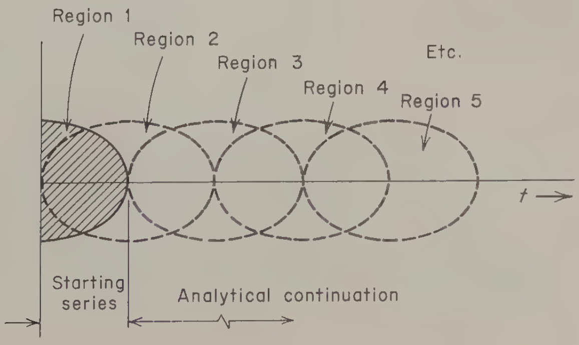


Figure 10.2

that an overlap with the first range exists? A greater degree of interplay between the polynomial representations of the function y in these two regions could then be facilitated, and the total problem [that is, $y(1.0 < t < 2.0)$] would separate to a consideration of the series representation of several overlapping intervals. It is to be noted, however, that the first region differs from the remaining ones in that special attention is required to obtain a sufficiently accurate *starting series. Analytical continuation* to the overlapping second, third, etc., regions is in general not as sensitive to round-off errors. This concept is illustrated in Fig. 10.3.

Several different methods can be used to achieve analytical continuation. The total range of independent variable in question can, for example, be orderly subdivided, and polynomials developed that exactly satisfy the differential equation at these preselected "pivotal points." Another approach is to select the undetermined coefficients of assumed generalized polynomials so that the total error over the entire range is a minimum. Still another criterion is to minimize the integral of the squares of the

Figure 10.3



errors. These methods for developing approximating polynomials are basically the same as the second general approach to the numerical solution of differential equations, referred to in Sec. 10.1. Analytical continuation is considered further in Sec. 10.8.

10.2.2 Sets of First-order Differential Equations

Sets of first-order ordinary differential equations can be solved by using a slightly modified form of the series method developed above. Consider, for example, the equations and boundary conditions defined by

$$\frac{dx}{dt} = 3x + y + t \qquad \frac{dy}{dt} = x - 2t \tag{10.15}$$

It is specified that at $t = 0$

$$x = 1.0 \qquad \text{and} \qquad y = -1.0$$

The derivatives with respect to t are as follows:

$$\begin{aligned}
&x(0) = 1.0 && y(0) = -1.0 \\
&\dot{x}(0) = 3x(0) + y(0) = 3 - 1 = +2.0 && \dot{y}(0) = x(0) = +1.0 \\
&\ddot{x}(0) = 3\dot{x}(0) + \dot{y}(0) + 1 = +8.0 && \ddot{y}(0) = \dot{x}(0) - 2 = 0 \\
&\dddot{x}(0) = 3\ddot{x}(0) + \ddot{y}(0) = +24.0 && \dddot{y}(0) = \ddot{x}(0) = +8.0 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

The Taylor series expansions for x and y about $t = 0$ are, therefore,

$$x(t) = x(0) + \dot{x}(0)t + \frac{\ddot{x}(0)t^2}{2!} + \frac{\dddot{x}(0)t^3}{3!} + \cdots$$

or

$$x(t) = 1.0 + 2t + 4t^2 + 4t^3 + \cdots \tag{10.16}$$

and

$$y(t) = y(0) + \dot{y}(0)t + \frac{\ddot{y}(0)t^2}{2!} + \frac{\dddot{y}(0)t^3}{3!} + \cdots$$

or

$$y(t) = -1.0 + t + \tfrac{4}{3}t^3 + \cdots \tag{10.17}$$

It is to be noted that any number of equations could be solved in this same manner. As is true for the single differential equation, the error in the solution depends both on the number of terms carried in the series approximation and on the magnitude of the interval of the independent variable t.

10.3 Series Determination by Iteration of Integral Equations: Initial-value Problems

For certain types of problems, where some of the successive, dependent derivatives become infinite, it may be impossible to obtain a Taylor series expansion in the manner described above. In other cases, the derivatives may be definable, but they may be too complicated or cumbersome to handle. In such cases, it may be more desirable to obtain an approximating series solution by successive iteration of the corresponding integral equation. Such a procedure is normally referred to as Picard's method.

10.3.1 First-order Differential Equations

Consider, for illustration of the general method, the first-order differential equation and boundary condition defined by

$$y' = A(x,y) \qquad y(x_0) = y_0 \tag{10.18}$$

The corresponding integral equation is

$$y = y_0 + \int_{x_0}^{x} A(x,y)\, dx \tag{10.19}$$

In the normal form for iteration this is

$${}_{k+1}y = y_0 + \int_{x_0}^{x} A(x,{}_ky)\, dx \tag{10.20}$$

where ${}_{k+1}y$ corresponds to the successive approximations to the solution.

As an example, consider the differential equation and initial condition defined by

$$y' = 1 - xy \qquad y(0) = 1.0 \tag{10.21}$$

The corresponding integral equation is

$$y = y_0 + \int_0^x (1 - xy)\, dx \tag{10.22}$$

or, in iteration form,

$${}_{k+1}y = 1.0 + \int_0^x (1 - x\,{}_ky)\, dx \tag{10.23}$$

By using as a starting value

$${}_1y = 1.0 \tag{10.24}$$

the successive iterations yield as approximate solutions

$$_2y = 1.0 + \int_0^x (1 - x)\,dx = 1.0 + x - \frac{x^2}{2} \tag{10.25}$$

$$_3y = 1.0 + \int_0^x \left[1 - x\left(1.0 + x - \frac{x^2}{2}\right)\right] dx = 1.0 + x - \frac{x^2}{2} - \frac{x^3}{3} + \frac{x^4}{8} \tag{10.26}$$

For comparison, the Taylor series expansion is shown in Eqs. (10.27) and (10.28).

$$\begin{aligned} y' &= 1 - xy & \text{or} \quad y'(0) & & &= +1 \\ y'' &= -y - xy' & y''(0) & & &= -1 \\ y''' &= -2y' - xy'' & y'''(0) &= -2(1) & &= -2 \\ y'''' &= -3y'' - xy''' & y''''(0) &= -3(-1) & &= +3 \end{aligned} \tag{10.27}$$

which gives

$$y = 1.0 + \frac{x}{1}(+1) + \frac{x^2}{2!}(-1) + \frac{x^3}{3!}(-2) + \frac{x^4}{4!}(+3) + \cdots$$

or $$y = 1.0 + x - \frac{x^2}{2} - \frac{x^3}{3} + \frac{x^4}{8} + \cdots \tag{10.28}$$

It should be noted that, as expected, both procedures give the identical series approximation.

10.3.2 Sets of First-order Differential Equations

Again, to facilitate comparison of the two methods, the equations listed as (10.15) are solved. They are here repeated as (10.29).

$$\begin{aligned} \frac{dx}{dt} &= 3x + y + t & \frac{dy}{dt} &= x - 2t \\ x(0) &= +1.0 & y(0) &= -1.0 \end{aligned} \tag{10.29}$$

The general integral equations are

$$_{k+1}x = x(0) + \int_0^t (3\,{}_kx + {}_ky + t)\,dt \tag{10.30}$$

$$_{k+1}y = y(0) + \int_0^t ({}_kx - 2t)\,dt \tag{10.31}$$

where $_{k+1}x$ and $_{k+1}y$ correspond to the successive approximations to the solution. These give as first approximations

$$_1x = x(0) + \int_0^t (3 - 1 + t)\,dt = 1 + 2t + \frac{t^2}{2} \tag{10.32}$$

and

$$_1y = y(0) + \int_0^t (1 - 2t)\,dt = -1 + t - t^2 \tag{10.33}$$

Second approximations are obtained in the following manner:

$$\begin{aligned} _2x &= x(0) + \int_0^t (3 + 6t + \tfrac{3}{2}t^2 - 1 + t - t^2 + t)\,dt \\ &= 1 + \int_0^t (2 + 8t + \tfrac{1}{2}t^2)\,dt \\ &= 1 + 2t + 4t^2 + \frac{t^3}{6} \end{aligned} \tag{10.34}$$

and

$$\begin{aligned} _2y &= y(0) + \int_0^t \left(1 + 2t + \frac{t^2}{2} - 2t\right) dt \\ &= -1 + t + \frac{t^3}{6} \end{aligned} \tag{10.35}$$

Equations (10.16) and (10.34) should be compared, as should (10.17) and (10.34). It is to be recognized, however, that (10.34) and (10.35) represent only second approximations. More nearly correct solutions are to be obtained by continuing the iteration process.

10.4 Polynomial Approximations

Basically, three approaches are used in the selection of general, analytical approximating functions. For example, general expressions containing a sufficient number of unknown coefficients can be chosen which identically satisfy the given differential equation. In this case the undetermined coefficients of these expressions are selected so that the boundary and/or initial conditions are best satisfied. A second approach is to prescribe a function that fits the boundary constraints and then, by one method or another, force it to satisfy (as well as can be) the differential equation. A third procedure is to select expressions which fit neither the differential equation nor the boundaries and then attempt to satisfy both. In the work that follows it is assumed that polynomial approximations are sufficiently general to allow solution by any, or all, of these three procedures.

In developing polynomial approximations, several points are to be kept in mind. It is not necessary to fit a single polynomial over the entire interval in question. As demonstrated in the section on Taylor series expansion, piecewise approximations may be more accurate since each approximation is required to represent only a part of the total interval. Piecewise approximations are also faster to evaluate, since smaller-degree polynomials can be used.

In all but the most unusual circumstances, polynomial approximations have errors associated with them. These deviations from the correct solution vary over the region in question, and the problem is to select the polynomial that results in the least deviation. The remainder of this section is concerned with various criteria that have been developed to minimize these errors. The four discussed are (1) collocation, (2) subdivision method, (3) least-squares method, and (4) Galerkin's method.

10.4.1 Collocation

This method, which is extensively developed and used later in this and in the following chapter, requires that the error vanish at certain preselected "pivotal points." For illustration of the procedure, consider again the differential equation and initial condition defined by (10.21). For convenience they are here repeated as Eqs. (10.36):

$$\frac{dy}{dx} = 1.0 - xy \qquad y(0) = 1.0 \tag{10.36}$$

It is assumed that the range of independent variable of interest is $0 \le x \le 0.75$.

The second-degree polynomial defined by Eq. (10.37), which contains two undetermined coefficients, is presumed as a sufficiently general approximating function. [It is to be noted that the chosen polynomial identically satisfies the initial condition $y(0) = 1.0$.]

$$y = 1.0 + a_1x + a_2x^2 \tag{10.37}$$

Substituting from Eq. (10.37) into (10.36) results in an expression in the unknown coefficients a_1 and a_2 and the independent variable x.

The error ϵ associated with this polynomial representation is a function of both the unknown coefficients a_1 and a_2 and the independent variable.

$$a_1(1 + x^2) + a_2(2x + x^3) - (1 - x) = \epsilon \tag{10.38}$$

Assume now that it is desired to make this error vanish at the two interior points $x = 0.25$ and 0.50. For this to be possible, the two linear equations (10.39) must be simultaneously satisfied.

$$\begin{aligned} &\text{At } x = 0.25: \qquad 1.063a_1 + 0.516a_2 = 0.750 \\ &\text{At } x = 0.50: \qquad 1.250a_1 + 1.125a_2 = 0.500 \end{aligned} \tag{10.39}$$

The solution to these equations is $a_1 = +1.063$ and $a_2 = -0.737$, and the corresponding approximating polynomial is

$$y = 1.0 + 1.063x - 0.737x^2 \tag{10.40}$$

For the case just considered, because of the nature of the assumed quadratic equation, collocation at two points was required. The two arbitrarily chosen pivotal points, $x = 0.25$ and 0.50, were within the interval in question. At these points the error term ϵ is made to vanish. It is to be noted, however, that while it is not forced by any algebraic manipulation, ϵ at $x = 0$ is also equal to zero. This is a direct consequence of the chosen polynomial.

10.4.2 Subdivision Method

The subdivision method presupposes that the total region in question is subdivided into as many smaller units as there are undetermined coefficients in the assumed polynomial. For the approximation defined earlier as Eq. (10.37), that is,

$$y = 1.0 + a_1x + a_2x^2 \tag{10.41}$$

there would be required two subdivisions. The undetermined coefficients are selected so that the integral of the error over each of these smaller intervals equals zero.

$$\int_X \epsilon \, dx = 0 \tag{10.42}$$

For the example, by assuming overlapping regions $0 < x < 0.5$ and $0.25 < x < 0.75$, the two equations to be satisfied are

$$\int_0^{0.50} \epsilon \, dx = 0 \qquad \int_{0.25}^{0.75} \epsilon \, dx = 0 \tag{10.43}$$

Substituting from the error equation (10.38) into each of these and integrating yields the following two linear equations in the unknown coefficients:

$$\begin{aligned} 0.542a_1 + 0.266a_2 &= 0.375 \\ 0.635a_1 + 0.578a_2 &= 0.250 \end{aligned} \tag{10.44}$$

The solution is $a_1 = 1.041$ and $a_2 = -0.711$. The corresponding approximating polynomial is

$$y = 1.0 + 1.041x - 0.711x^2 \tag{10.45}$$

10.4.3 Least-squares Method

The least-squares method requires that the integral of the square error over the entire range of x in question be made as small as possible. By defining

$$J = \text{mean-square error} = \int_X \epsilon^2\,dx \tag{10.46}$$

the necessary conditions that J be a minimum are

$$\frac{\partial J}{\partial a_i} = \frac{\partial}{\partial a_i}\int_X \epsilon^2\,dx = \int_X \epsilon\frac{\partial \epsilon}{\partial a_i}\,dx = 0 \tag{10.47}$$

An nth-degree polynomial therefore requires the evaluation of n separate integrations.

Again consider for illustration of the general method of solution the example considered in the preceding two cases. The error is defined by Eq. (10.38), here repeated as (10.48).

$$\epsilon = a_1(1 + x^2) + a_2(2x + x^3) - (1 - x) \tag{10.48}$$

The partial derivatives are

$$\frac{\partial \epsilon}{\partial a_1} = 1 + x^2 \qquad \frac{\partial \epsilon}{\partial a_2} = 2x + x^3 \tag{10.49}$$

The general integrals to be evaluated are

$$\int_0^{0.75} [a_1(1 + x^2) + a_2(2x + x^3) - (1 - x)](1 + x^2)\,dx = 0 \tag{10.50}$$

$$\int_0^{0.75} [a_1(1 + x^2) + a_2(2x + x^3) - (1 - x)](2x + x^3)\,dx = 0 \tag{10.51}$$

The resulting linear algebraic equations in a_1 and a_2 are those given as (10.52). The solution is $a_1 = 1.041$ and $a_2 = -0.715$.

$$\begin{aligned} 1.079a_1 + 0.830a_2 &= 0.530 \\ 0.830a_1 + 0.771a_2 &= 0.313 \end{aligned} \tag{10.52}$$

The corresponding approximating polynomial is

$$y = 1.0 + 1.041x - 0.715x^2 \tag{10.53}$$

It should be noted that in certain instances weighting functions are used with the least-squares method. If these functions are denoted as W_i, solution requires that

$$J = \int_X W_i\epsilon^2\,dx = \text{minimum} \tag{10.54}$$

10.4.4 Galerkin's Method

Galerkin's method, which is a special case of the more general orthogonality method, entails that the error is orthogonal over the range in question to a set of linearly independent weighting functions ϕ_i of the independent variable. That is,

$$\int_X \phi_i \epsilon \, dx = 0 \tag{10.55}$$

Commonly in Galerkin's method the ϕ_i terms are taken as the individual elements of the original trial polynomial. For the example problem, $\phi_1 = x$ and $\phi_2 = x^2$. If the original polynomial were of the form

$$y = 1.0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n \tag{10.56}$$

n separate integrals over the entire interval would need to be evaluated.

$$\int_X \epsilon x \, dx = \int_X \epsilon x^2 \, dx = \int_X \epsilon x^3 \, dx = \cdots = \int_X \epsilon x^n \, dx = 0 \tag{10.57}$$

Again, for the example under consideration [Eqs. (10.36) and (10.37)], the integrals are

$$\int_0^{0.75} x[a_1(1 + x^2) + a_2(2x + x^3) - (1 - x)] \, dx = 0 \tag{10.58}$$

$$\int_0^{0.75} x^2[a_1(1 + x^2) + a_2(2x + x^3) - (1 - x)] \, dx = 0 \tag{10.59}$$

Evaluating these integrals gives the set of equations to be solved as

$$0.3604a_1 + 0.3287a_2 = 0.1407 \tag{10.60}$$

$$0.1881a_1 + 0.1879a_2 = 0.0615 \tag{10.61}$$

Its solution is $a_1 = 1.055$ and $a_2 = -0.729$. The approximating polynomial is therefore

$$y = 1.0 + 1.055x - 0.729x^2 \tag{10.62}$$

Table 10.2 is a summary of the polynomial approximations that have been obtained in this section. The corresponding Taylor series is also given in the table.

10.5 Higher-order Differential Equations: Initial-value Problems

While it is possible to develop special techniques to facilitate the direct solution of certain types of higher-order differential equations, it is also possible to reduce higher-order equations to sets of first-order ones, and thereby allow solution by the methods hereinbefore described. Normally, this reduction takes the form of substitution of variables in which the nth-order equation is replaced by a series of n first-order equations. The following three examples illustrate the general procedure.

Table 10.2

Given equation: $\dfrac{dy}{dx} = 1.0 - xy \qquad 0 \le x \le 0.75$

Given initial condition: $y(0) = 1.0$

		Function evaluated at selected points.			
Method	Approximating polynomial	$y(0)$	$y(0.25)$	$y(0.50)$	$y(0.75)$
Collocation	$y = 1.0 + 1.063x - 0.737x^2$	1.000	1.220	1.347	1.384
Subdivision	$y = 1.0 + 1.041x - 0.711x^2$	1.000	1.216	1.343	1.381
Least squares	$y = 1.0 + 1.041x - 0.715x^2$	1.000	1.215	1.342	1.379
Galerkin	$y = 1.0 + 1.055x - 0.729x^2$	1.000	1.218	1.346	1.381
Taylor series	$\left\{ y = 1.0 + x - \frac{x^2}{2} - \frac{x^3}{3} + \frac{x^4}{8} + \frac{x^5}{15} + \frac{x^6}{48} - \frac{x^7}{105} - \cdots \right\}$	1.000	1.214	1.344	1.387

Example 10.1 Given the differential equation

$$\frac{d^2y}{dx^2} + x\frac{dy}{dx} + y^2 + c = 0 \tag{10.63}$$

If a new variable z is defined such that

$$z = \frac{dy}{dx} \tag{10.64}$$

then

$$\frac{dz}{dx} = \frac{d^2y}{dx^2} \tag{10.65}$$

and Eq. (10.63) becomes

$$\frac{dz}{dx} + xz + y^2 + c = 0 \tag{10.66}$$

The second required equation to allow a solution is given by (10.64), here repeated as (10.67).

$$\frac{dy}{dx} - z = 0 \tag{10.67}$$

Example 10.2 Given the set of equations

$$\frac{dx}{dt} = 3x + y + t \qquad \frac{d^2y}{dt^2} = x - 2t \tag{10.68}$$

The initial conditions are

$$x(0) = 0 \qquad y(0) = -1 \qquad \dot{y}(0) = 0 \tag{10.69}$$

If it is assumed that

$$z = \frac{dy}{dt} \tag{10.70}$$

then, from the second of Eqs. (10.68),

$$\frac{dz}{dt} = x - 2t \tag{10.71}$$

The corresponding set of first-order equations to be solved is

$$\begin{aligned} \frac{dx}{dt} &= 3x + y + t \\ \frac{dy}{dt} &= z \\ \frac{dz}{dt} &= x - 2t \end{aligned} \tag{10.72}$$

where $$x(0) = 0 \qquad y(0) = -1 \qquad z(0) = 0 \tag{10.73}$$

Example 10.3

$$\frac{d^4y}{dx^4} - \lambda^4 \frac{d^2y}{dx^2} + ky = 0 \tag{10.74}$$

where
$$\begin{aligned} y(0) &= y_0 \\ y'(0) &= y_0' \\ y''(0) &= y_0'' \\ y'''(0) &= y_0''' \end{aligned} \tag{10.75}$$

By selecting

$$A = \frac{dy}{dx}$$

$$B = \frac{dA}{dx} = \frac{d^2y}{dx^2}$$

$$C = \frac{dB}{dx} = \frac{d^3y}{dx^3}$$

Eq. (10.74) reduces to

$$\frac{dy}{dx} = A \qquad \frac{dA}{dx} = B$$
$$\frac{dB}{dx} = C \qquad \frac{dC}{dx} = \lambda^4 B - ky \tag{10.76}$$

where

$$y(0) = y_0$$
$$A(0) = y_0'$$
$$B(0) = y_0'' \tag{10.77}$$
$$C(0) = y_0'''$$

10.6 Euler Methods of Forward Integration

Many methods for the solution of the general first-order differential equation

$$\frac{dy}{dx} = f(x,y) \tag{10.78}$$

are based on the equation

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(x,y)\,dx \tag{10.79}$$

Often they are written in the form

$$\Delta y = y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} f(x,y)\,dx \tag{10.80}$$

This section is a consideration of the first, and most crude, of these, the Euler method. It is not recommended for use except under the most unusual circumstances, where accuracy is of secondary importance. It is presented here for the purpose of indicating a general approach to the solution of a problem. Also, it does allow modification, as is discussed, and facilitates the introduction to the more usable Runge-Kutta methods of forward integration, which are presented in detail in Sec. 10.7.

10.6.1 Euler's Method

The basic presumption of the Euler method is that the function $f(x,y)$ can be assumed to be constant at its initial value $f(x_i,y_i)$ over the interval of integration $x_i \leq x \leq x_{i+1}$. That is,

$$\Delta y = y_{i+1} - y_i = y_{x+h} - y_x = k_1 \tag{10.81}$$

where $$k_1 = h[f(x_i,y_i)] \tag{10.82}$$

This can equally well be written

$$\Delta y = y(x + h) - y(x) = k_1 \tag{10.83}$$

$$k_1 = h[f(x,y)] \tag{10.84}$$

The results of this type of reasoning can be seen from Fig. 10.4. If $y''(x) > 0$, then the Euler method yields a value for $y(x + h)$ which is less than the correct value. For $y''(x) < 0$, the value obtained is too great.

In terms of overall general accuracy of solution, an improvement in the above-defined method is realized if the function is considered at some intermediate point $x_i \leq x \leq x_{i+1}$; for example, at the midpoint of the interval. The governing equations are

$$\Delta y = k_1 \tag{10.85}$$

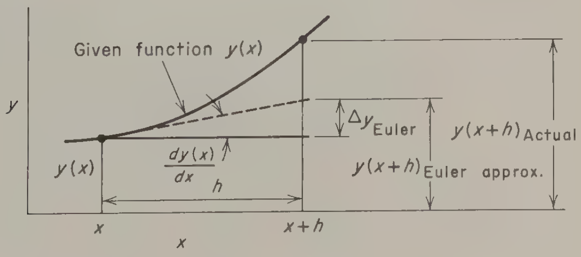
and $$k_1 = h\{f[x + \tfrac{1}{2}h,\ y(x + \tfrac{1}{2}h)]\} \tag{10.86}$$

The major difficulty associated with this improvement is emphasized by the brackets and parentheses contained in Eq. (10.86). y, which is needed to establish the function value f, is a function of the independent variable x. Therefore, f is not readily defined.

10.6.2 Modified Euler-Cauchy Method

Another method for improving the Euler method can be realized by averaging the slopes of the function $y(x)$ at its end points, and by presuming that this average value is constant over the entire interval h. Unfortunately, the same problem exists as discussed in the preceding case. The function value, and therefore the slope of the function at $x + h$, is not readily available. Approximations to it can, however, be established.

Figure 10.4



In the modified Euler-Cauchy method it is assumed that

$$\Delta y = \frac{h}{2}\left[f(x_i,y_i) + \tilde{f}(x_{i+1},y_{i+1})\right] \tag{10.87}$$

where $\tilde{f}(x_{i+1},y_{i+1})$ is an approximation to the true value of the function at $x + h$. It is further presumed that the Euler method [Eqs. (10.83) and (10.84)] is sufficiently accurate to establish this approximate function value. In the notation of the preceding case, the modified Euler-Cauchy method is defined by Eqs. (10.88) and (10.89).

$$\Delta y = \tfrac{1}{2}(k_1 + k_2) \tag{10.88}$$

$$\begin{aligned} k_1 &= h[f(x,y)] \\ k_2 &= h[f(x + h,\ y + k_1)] \end{aligned} \tag{10.89}$$

As an example of the numerical procedure that results in the solution of problems by use of this method, consider the evaluation of $y(0.5)$, given the differential equation

$$y' = f(x,y) = 1 - xy \tag{10.90}$$

It is specified that the initial condition is $y(0) = 1.0$. For a "one-step" forward integration the interval h would be selected as the full value of the independent variable. That is, $h = 0.5$. From the first of equations (10.89),

$$f(x,y) = f(0,1.0) = 1 - (0)(1.0) = 1.0$$

Therefore

$$k_1 = h[f(x,y)] = (0.5)(1.0) = 0.5 \tag{10.91}$$

The second of Eqs. (10.89) is evaluated in the following sequence:

$$x + h = 0 + 0.5 = 0.5$$

$$y + k_1 = 1.0 + 0.5 = 1.5$$

and

$$\begin{aligned} f(x + h,\ y + k_1) = f(0.5,1.5) &= 1 - xy \\ &= 1 - (0.5)(1.5) \\ &= 0.25 \end{aligned}$$

Therefore

$$\begin{aligned} k_2 &= h[f(x + h,\ y + k_1)] \\ &= (0.5)[f(0.5,1.5)] = (0.5)(0.25) = 0.125 \end{aligned} \tag{10.92}$$

From Eq. (10.88)

$$\Delta y = \tfrac{1}{2}(k_1 + k_2) = \tfrac{1}{2}(0.5 + 0.125) = 0.3125 \tag{10.93}$$

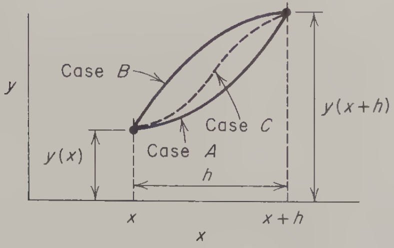


Figure 10.5

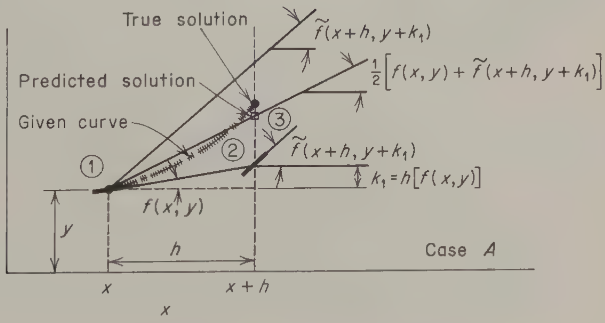
This gives as a value of the function $y(0.5)$

$$y(0.5) = y(0) + \Delta y = 1.0 + 0.3125 = 1.3125 \qquad (10.94)$$

It is evident from the above that the methods presented in this section result in an accuracy of solution which depends primarily on the selected interval h. Accuracy, however, also depends on the overall behavior of the function in question, as is now demonstrated. Consider the three cases shown in Fig. 10.5. In each it is presumed that the function starts at x, $y(x)$ and ends at $x + h$, $y(x + h)$. In case A it is assumed that in the interval in question $y''(x) > 0$. For case B, the presumption is that $y''(x) < 0$. Case C presupposes a point of inflection within the interval.

Figure 10.6 illustrates graphically the procedure that is followed in obtaining an approximate value for $y(x + h)$ by using the Euler-Cauchy method. The initial values x and y are presumed given. On the figure they are shown as location ①. With these starting values and the given first-order differential equation, it is possible to establish the slope of the function at ①. That is, $y' = f(x,y)$. A first approximation to the value of y at $x + h$ can be realized by assuming, as in the Euler method, that y' remains constant at its initial value, over the interval h. (This is

Figure 10.6



shown as location ② on the figure.) With this first approximation to the value of $y(x + h)$, an approximate slope at $x + h$ can now be determined, $\tilde{f}(x + h, y + k_1)$. A second approximation to the true solution $y(x + h)$ can be obtained by averaging the slopes at ① and ② (that is, weighing them equally) and using this average value over the entire interval. Location ③ would be the resulting predicted solution. It should be evident that for case *A* type curves $[y''(x) > 0]$, location ③ is always a closer approximation to the true solution than is location ②.

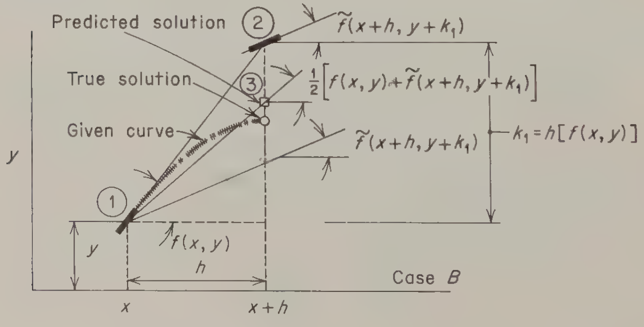
The Euler-Cauchy method applied to case *B* type curves is illustrated in Fig. 10.7. Again, location ③ is a better approximation than is location ②.

For case *C*, where $y''(x) = 0$ at some point within the interval in question, it is not possible to reason that location ③ is always a better approximation to the true solution than is location ②. Moreover, if the slopes at ① and ② are of opposite signs, ② may be appreciably better than ③.

There are several ways to improve the accuracy of solution of the methods developed in this section. For example, the interval h could be subdivided into several smaller regions and there could be developed algebraic equations which, in turn, apply the Euler-Cauchy method to each subdivision and sum the resulting values. However, very little, if any, computational advantage would be realized by such a procedure over just selecting a smaller interval h and repeating the process the desired number of times. Such a method is not developed here.

From Figs. 10.6 to 10.8 it is evident that the slopes of the functions at the terminal points of the interval in question may not be necessarily the best indicator of the "effective average slope" over the interval itself. A second method for improving the accuracy is therefore suggested: a weighted averaging of approximate slopes at several different locations. Assume for illustration of the general method that the predicted change in y over the interval h (that is, Δy) can be defined in terms of approximate (or exact) slopes at four different locations.

Figure 10.7



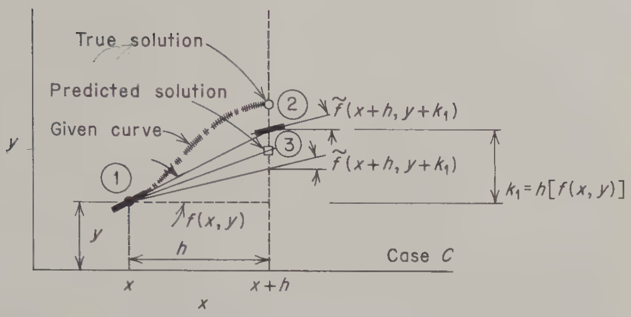


Figure 10.8

$$\Delta y = ak_1 + bk_2 + ck_3 + dk_4 \tag{10.95}$$

where a, b, c, and d are weighting values and k_1, k_2, k_3, and k_4 are h times the slopes at four points within the interval in question. If it is assumed that all the slopes are to contribute in a positive fashion, the weighting functions must be 1, or less than 1, and must sum to 1.0. That is,

$$a + b + c + d = 1 \tag{10.96}$$

and

$$a \leq 1.0 \qquad b \leq 1.0 \qquad c \leq 1.0 \qquad d \leq 1.0 \tag{10.97}$$

The computed approximate slopes used in determining k_1 through k_4 could be defined in one of several ways. For example, each could be prescribed using the Euler method, assuming as a starting point the initial condition x, y.

$$\begin{aligned} k_1 &= h[f(x,y)] \\ k_2 &= h[f(x + mh, y + mk_1)] \\ k_3 &= h[f(x + nh, y + nk_1)] \\ k_4 &= h[f(x + ph, y + pk_1)] \end{aligned} \tag{10.98}$$

However, if it is specified that $m \leq n \leq p \leq 1.0$, possibly a more realistic solution would be realized by using the following definition of terms:

$$\begin{aligned} k_1 &= h[f(x,y)] \\ k_2 &= h[f(x + mh, y + mk_1)] \\ k_3 &= h[f(x + nh, y + nk_2)] \\ k_4 &= h[f(x + ph, y + pk_3)] \end{aligned} \tag{10.99}$$

Other definitions could equally be specified, the accuracy of the solution depending not only on these equations but also on the selected weighting functions a through d.

10.7 *Runge-Kutta Methods*

10.7.1 First-order Equations

Higher-order approximations of the basic forward-integration type discussed in Sec. 10.6 can be developed in a more general form by using the methods of Runge-Kutta. These methods provide solutions which are comparable in accuracy to Taylor series solutions in which are retained many higher-order derivatives. They are much used for accurate integration.

Again, it is presumed that the differential equation in question is of the first order and can be expressed in the following form:

$$\frac{dy}{dx} = f(x,y) \tag{10.100}$$

It should be noted that this equation need not be linear. A Taylor series expansion about point x, y of the function $y(x)$ is

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2!}y''(x) + \frac{h^3}{3!}y'''(x) + \cdots \tag{10.101}$$

where h is the interval of expansion in question. This equation can equally well be written in terms of the change in y, or Δy, associated with a change in x from x to $x + h$.

$$\Delta y = y(x + h) - y(x) = hy'(x) + \frac{h^2}{2!}y''(x) + \frac{h^3}{3!}y'''(x) + \cdots \tag{10.102}$$

It is to be recalled that the Taylor series method for obtaining starting solutions allows the definition of higher-order dependent derivatives. For the given differential equation (10.100), the second derivative of function $y(x)$ with respect to x is

$$y''(x) = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}\frac{dy}{dx} \tag{10.103}$$

But from the differential equation, $dy/dx = f(x,y)$. Therefore, Eq. (10.103) can be written as

$$y''(x) = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}f(x,y) \tag{10.104}$$

For compactness, the following notation is used in the development of the Runge-Kutta methods:

$$\begin{array}{lll} f(x,y) = f & & \\ \dfrac{\partial f(x,y)}{\partial x} = f_x & \dfrac{\partial f_x}{\partial x} = f_{xx} & \dfrac{\partial f_{xx}}{\partial x} = f_{xxx} \\ \dfrac{\partial f(x,y)}{\partial y} = f_y & \dfrac{\partial f_y}{\partial y} = f_{yy} & \dfrac{\partial f_{yy}}{\partial y} = f_{yyy} \\ & \dfrac{\partial f_x}{\partial y} = f_{xy} & \dfrac{\partial f_{xy}}{\partial y} = f_{xyy} \\ & \cdots\cdots\cdots & \cdots\cdots\cdots \end{array} \tag{10.105}$$

The derivatives can then be expressed as follows:

$$y' = f \tag{10.106}$$

$$y'' = f_x + f_y f \tag{10.107}$$

$$y''' = f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f) \tag{10.108}$$

$$\begin{aligned} y'''' = f_{xxx} &+ 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3 + f_y(f_{xx} + 2f_{xy}f + f_{yy}f^2) \\ &+ 3(f_x + f_y f)(f_{xy} + f_{yy}f) + f_y{}^2(f_x + f_y f) \end{aligned} \tag{10.109}$$

In turn, Eq. (10.102), the Taylor series, will be

$$\begin{aligned} \Delta y = hf &+ \frac{h^2}{2}(f_x + f_y f) + \frac{h^3}{6}[f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f)] \\ &+ \frac{h^4}{24}[f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3 + f_y(f_{xx} + 2f_{xy}f + f_{yy}f^2) \\ &\qquad + 3(f_x + f_y f)(f_{xy} + f_{yy}f) + f_y{}^2(f_x + f_y f)] + \cdots \end{aligned} \tag{10.110}$$

The Runge-Kutta methods provide a straightforward means for developing expressions for Δy in terms of values of the function $f(x,y)$ at various locations along the interval in question. Moreover, the resulting formulas satisfy Eq. (10.110), the Taylor series, to the accuracy desired.

As in Eq. (10.95), it is assumed that Δy can be expressed in the following form:

$$\Delta y = ak_1 + bk_2 + ck_3 + dk_4 \tag{10.111}$$

where, by definition, as in Eq. (10.99),

$$
\begin{aligned}
k_1 &= h[f(x,y)] \\
k_2 &= h[f(x + mh,\ y + mk_1)] \\
k_3 &= h[f(x + nh,\ y + nk_2)] \\
k_4 &= h[f(x + ph,\ y + pk_3)]
\end{aligned}
\tag{10.112}
$$

Each of these equations (10.112) can be expanded in a Taylor series in two independent variables. That is,

$$k_i = h[f(x + \alpha,\ y + \beta)] \tag{10.113}$$

where

$$
\begin{aligned}
f(x + \alpha,\ y + \beta) &= f(x,y) + \alpha f_x(x,y) + \beta f_y(x,y) \\
&\quad + \frac{1}{2!}[\alpha^2 f_{xx}(x,y) + 2\alpha\beta f_{xy}(x,y) + \beta^2 f_{yy}(x,y)] + \cdots
\end{aligned}
\tag{10.114}
$$

By using the notation of Eqs. (10.105), this is written as

$$f(x + \alpha,\ y + \beta) = f + \alpha f_x + \beta f_y + \frac{1}{2!}(\alpha^2 f_{xx} + 2\alpha\beta f_{xy} + \beta^2 f_{yy}) + \cdots \tag{10.115}$$

The corresponding expression for k_2 is

$$\alpha = mh \tag{10.116}$$

$$\beta = mk_1 = mhf \tag{10.117}$$

which gives

$$
\begin{aligned}
k_2 = h[f + mhf_x + mk_1 f_y + \tfrac{1}{2}(m^2h^2 f_{xx} + 2mhmk_1 f_{xy} + m^2k_1^2 f_{yy}) \\
+ \cdots]
\end{aligned}
\tag{10.118}
$$

or

$$
\begin{aligned}
k_2 = h\Bigg[f + mh(f_x + f_y f) + \frac{m^2h^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) \\
+ \frac{m^3h^3}{6}(f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3) + \cdots\Bigg]
\end{aligned}
\tag{10.119}
$$

Expressions for k_3 and k_4 are to be obtained in a similar manner. The resulting expressions are listed in Table 10.3.

It is the primary purpose of the Runge-Kutta methods to establish the interrelationships between the constants a, b, c, d, m, n, and p such that Eqs. (10.110) and (10.111) correspond to the degree of accuracy desired. This could be done in a number of ways. The method selected is to have the coefficients of h, h^2, h^3, and h^4 in both systems equal. For example, the coefficient of h in Eq. (10.110) is f. The term in the expanded series corresponding to Eq. (10.111) is $af + bf + cf + df$. It is therefore required that

$$f = af + bf + cf + df$$

or

$$a + b + c + d = 1.0 \tag{10.120}$$

In like manner, matching the coefficients of h^2 in each system,

$$\tfrac{1}{2}(f_x + f_y f) = bm(f_x + f_y f) + cn(f_x + f_y f) + dp(f_x + f_y f) \tag{10.121}$$

or

$$bm + cn + dp = \tfrac{1}{2} \tag{10.122}$$

Table 10.3 Summary of Coefficients

Multipliers → k_i ↓	h	h^2	h^3	h^4
$k_1 =$	f	0	0	0
$k_2 =$	f	$m(f_x + f_y f)$	$\frac{m^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2)$	$\frac{m^3}{6}(f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3)$
$k_3 =$	f	$n(f_x + f_y f)$	$\frac{n^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) + mnf_y(f_x + f_y f)$	$\frac{n^3}{6}(f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3) + \tfrac{1}{2}m^2nf_y(f_{xx} + 2f_{xy}f + f_{yy}f^2) + mn^2(f_{xy} + f_{yy}f)(f_x + f_y f)$
$k_4 =$	f	$p(f_x + f_y f)$	$\frac{p^2}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) + npf_y(f_x + f_y f)$	$\frac{p^3}{6}(f_{xxx} + 3f_{xxy}f + 3f_{xyy}f^2 + f_{yyy}f^3) + \tfrac{1}{2}n^2pf_y(f_{xx} + 2f_{xy}f + f_{yy}f^2) + np^2(f_{xy} + f_{yy}f)(f_x + f_y f) + mnpf_y^2(f_x + f_y f)$

For the coefficients of h^3, it is presupposed that the necessary conditions can be effected in two separate steps, as is now demonstrated. The total equation is

$$\tfrac{1}{6}(f_{xx} + 2f_{xy}f + f_{yy}f^2) + \tfrac{1}{6}f_y(f_x + f_y f) = \tfrac{1}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2) \times (m^2b + n^2c + p^2d) + [f_y(f_x + f_y f)](cmn + dnp) \tag{10.123}$$

This equation, however, can be satisfied by subdividing it into the following separate equations:

$$\tfrac{1}{6}(f_{xx} + 2f_{xy}f + f_{yy}f^2) = \tfrac{1}{2}(f_{xx} + 2f_{xy}f + f_{yy}f^2)(m^2b + n^2c + p^2d) \tag{10.124}$$

and $$\tfrac{1}{6}[f_y(f_x + f_y f)] = [f_y(f_x + f_y f)](cmn + dnp) \qquad (10.125)$$

These give Eqs. (10.126) and (10.127).

$$m^2b + n^2c + p^2d = \tfrac{1}{3} \qquad (10.126)$$

$$cmn + dnp = \tfrac{1}{6} \qquad (10.127)$$

Table 10.4 summarizes the "sufficient" interrelationship, developed in the manner just described, that must exist to effect a solution having an accuracy up to the order of h^5. Any set of values satisfying these makes it possible to obtain a solution.

Table 10.4 Sufficient Conditions for Runge-Kutta Methods

Order of accuracy	Equations for determining coefficients
h	$a + b + c + d = 1$
h^2	$bm + cn + dp = \frac{1}{2}$
h^3	$bm^2 + cn^2 + dp^2 = \frac{1}{3}$ $cmn + dnp = \frac{1}{6}$
h^4	$bm^3 + cn^3 + dp^3 = \frac{1}{4}$ $cmn^2 + dnp^2 = \frac{1}{8}$ $cm^2n + dn^2p = \frac{1}{12}$ $dmnp = \frac{1}{24}$

As an example of how the information contained in Table 10.4 might be used to develop a set of Runge-Kutta type equations, assume that it is desired to develop an expression involving only two terms, which has an accuracy of the order of h^3. That is, the desired equation should be of the form

$$\Delta y = ak_1 + bk_2 \qquad (10.128)$$

with $$k_1 = h[f(x,y)] \qquad k_2 = h[f(x + mh,\, y + mk_1)] \qquad (10.129)$$

Using the equations hereinbefore developed, it is required that

$$c = d = n = p = 0 \qquad (10.130)$$

The required expressions, from Table 10.4, are

$$a + b = 1 \qquad (10.131)$$

$$bm = \tfrac{1}{2} \qquad (10.132)$$

Selecting $m = 1$ yields values for a and b as

$$m = 1 \qquad a = \tfrac{1}{2} \qquad b = \tfrac{1}{2} \tag{10.133}$$

It is evident that such a selection corresponds to the Euler-Cauchy method defined in Sec. 10.6.2. [It is to be recognized also that this is not a unique set for the three-variable two-equation system of (10.131) and (10.132).]

A Runge-Kutta set of equations which is accurate to the order h^3, and which presumes the following values of the constants,

$$\begin{aligned} a &= \tfrac{1}{4} & m &= \tfrac{1}{3} \\ b &= 0 & n &= \tfrac{2}{3} \\ c &= \tfrac{3}{4} & p &= 0 \\ d &= 0 \end{aligned} \tag{10.134}$$

satisfies the "sufficient" equations listed in Table 10.4. The resulting set of expressions, known as Heun's method, is

$$\begin{aligned} \Delta y &= \tfrac{1}{4}(k_1 + 3k_3) \\ k_1 &= h[f(x,y)] \\ k_2 &= h[f(x + \tfrac{1}{3}h,\ y + \tfrac{1}{3}k_1)] \\ k_3 &= h[f(x + \tfrac{2}{3}h,\ y + \tfrac{2}{3}k_2)] \end{aligned} \tag{10.135}$$

Several of the more commonly used Runge-Kutta methods are given in Table 10.5.

To give an indication of the accuracy of these various methods, and to compare them with the corresponding Taylor series expansion, consider the numerical solutions of the differential equation

$$y' = 1 - xy \tag{10.136}$$

As before, it is specified for all cases that $y(0) = 1.0$. The results are given in Table 10.6.

It is pointed out above in this section that, like most other numerical procedures for solving differential equations, the methods of Runge-Kutta are sensitive to variations in the selected interval h. Too small a value leads to an excessive number of computations, with many digits retained. On the other hand, too large an interval may well lead to an answer that differs significantly from the true one.

A method often used to examine the accuracy of a solution is to first solve the given problem for a trial selected interval h and then repeat the process, through two cycles, for an interval $\tfrac{1}{2}h$. It should be noted that such a procedure has the added advantage of providing a "current" check on the numerical calculations. As an example of the influence of "halv-

Table 10.5 Runge–Kutta-type Equations for $y' = f(x,y)$, $y(x + h) = y(x) + \Delta y$

Method	Equations
Euler (Error of the order of h^2)	$\Delta y = k_1$ $k_1 = h[f(x,y)]$
Euler-Cauchy (Error of the order of h^3)	$\Delta y = \frac{1}{2}(k_1 + k_2)$ $k_1 = h[f(x,y)]$ $k_2 = h[f(x + h,\ y + k_1)]$
Heun (Error of the order of h^4)	$\Delta y = \frac{1}{4}(k_1 + 3k_3)$ $k_1 = h[f(x,y)]$ $k_2 = h[f(x + \frac{1}{3}h,\ y + \frac{1}{3}k_1)]$ $k_3 = h[f(x + \frac{2}{3}h,\ y + \frac{2}{3}k_2)]$
Kutta-Simpson (One-third rule) (Error of the order of h^5)	$\Delta y = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$ $k_1 = h[f(x,y)]$ $k_2 = h[f(x + \frac{1}{2}h,\ y + \frac{1}{2}k_1)]$ $k_3 = h[f(x + \frac{1}{2}h,\ y + \frac{1}{2}k_2)]$ $k_4 = h[f(x + h,\ y + k_3)]$
Kutta-Simpson (Three-eighths rule) (Error of the order of h^5)	$\Delta y = \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4)$ $k_1 = h[f(x,y)]$ $k_2 = h[f(x + \frac{1}{3}h,\ y + \frac{1}{3}k_1)]$ $k_3 = h[f(x + \frac{2}{3}h,\ y + k_2 - \frac{1}{3}k_1)]$ $k_4 = h[f(x + h,\ y + k_3 - k_2 + k_1)]$

ing the interval," again consider the first-order differential equation

$$y' = 1 - xy \tag{10.137}$$

with

$$y(0) = +1.0 \tag{10.138}$$

The Kutta-Simpson one-third rule equations are used for illustration. If h is selected equal to 0.25 and if expansion about $x = 0$ is presumed, the necessary terms in the solution are

$$k_1 = 0.25[f(0,1.0)] = 0.25$$

$$k_2 = 0.25[f(0.125,1.125)] = 0.214844$$

$$k_3 = 0.25[f(0.125,1.107285)] = 0.215393$$

$$k_4 = 0.25[f(0.25,1.215397)] = 0.174038$$

$$\Delta y = \tfrac{1}{6}(0.25 + 0.429688 + 0.430786 + 0.174038) = 0.214085$$

and

$$y(0.25) = y(0) + \Delta y = 1.00 + 0.214085 = 1.214085$$

Table 10.6

		Function evaluated for selected intervals of expansion			
Method	General equations	$y(0)$	$y(0.25)$	$y(0.50)$	$y(0.75)$
Euler	$\Delta y = k_1$	1.000	1.250	1.500	1.750
Euler-Cauchy	$\Delta y = \frac{1}{2}(k_1 + k_2)$	1.000	1.212	1.313	1.258
Heun	$\Delta y = \frac{1}{4}(k_1 + 3k_3)$	1.000	1.212	1.342	1.372
Kutta-Simpson (One-third rule)	$\Delta y = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$	1.000	1.214	1.343	1.390
Kutta-Simpson (Three-eighths rule)	$\Delta y = \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4)$	1.000	1.214	1.343	1.385
Taylor series solution	$\Delta y = x - \frac{x^2}{2} - \frac{x^3}{3} + \frac{x^4}{8} + \frac{x^5}{15} + \frac{x^6}{48} - \frac{x^7}{105} - \cdots$	1.000	1.214	1.344	1.387

Table 10.7 summarizes this and other solutions. Note the comparable function values, using both a one-step application of the Runge-Kutta equations and a two-step application.

Table 10.7 $dy/dx = 1 - xy$, $y(0) = +1.0$

Selected interval of expansion h	Function values			
	$y(0.25)$	$y(0.50)$	$y(1.00)$	$y(2.00)$
0.25	1.213991	1.342720		
0.50		1.34395	1.31824	
1.00			1.3229	+0.7474
2.00				−0.3333

10.7.2 Sets of First-order Equations

Forward integration methods of the Runge-Kutta type can be developed to facilitate solution of sets of first-order differential equations. For two equations the relationships summarized in Table 10.8 are often used. By comparing these with that developed for single equations (see Table 10.5), it is evident that the method can be extended to larger numbers of equations.

Table 10.8

Given differential equation $\dfrac{dx}{dt} = f(t,x,y)$	*Given differential equation* $\dfrac{dy}{dt} = g(t,x,y)$
$k_1 = h[f(t,x,y)]$	$\ell_1 = h[g(t,x,y)]$
$k_2 = h[f(t + \frac{1}{2}h,\ x + \frac{1}{2}k_1,\ y + \frac{1}{2}\ell_1)]$	$\ell_2 = h[g(t + \frac{1}{2}h,\ x + \frac{1}{2}k_1,\ y + \frac{1}{2}\ell_1)]$
$k_3 = h[f(t + \frac{1}{2}h,\ x + \frac{1}{2}k_2,\ y + \frac{1}{2}\ell_2)]$	$\ell_3 = h[g(t + \frac{1}{2}h,\ x + \frac{1}{2}k_2,\ y + \frac{1}{2}\ell_2)]$
$k_4 = h[f(t + h,\ x + k_3,\ y + \ell_3)]$	$\ell_4 = h[g(t + h,\ x + k_3,\ y + \ell_3)]$
$\Delta x = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$	$\Delta y = \frac{1}{6}(\ell_1 + 2\ell_2 + 2\ell_3 + \ell_4)$
$x(t + h) = x(t) + \Delta x$	$y(t + h) = y(t) + \Delta y$

Consider, for illustration, the solution of the two first-order equations shown as (10.139). [These are the same equations that were solved earlier by using the Taylor series. See Eqs. (10.15).]

$$\frac{dx}{dt} = t + 3x + y \qquad \frac{dy}{dt} = -2t + x \tag{10.139}$$

where $x(0) = +1.0$ and $y(0) = -1.0$

If $h = 0.2$ is selected, solutions are as follows:

$k_1 = 0.2[f(0,1.0,-1.0)] = +0.400$

$\ell_1 = 0.2[g(0,1.0,-1.0)] = +0.200$

$k_2 = 0.2[f(0.1,1.2,-0.9)] = +0.560$

$\ell_2 = 0.2[g(0.1,1.2,-0.9)] = +0.200$

$k_3 = 0.2[f(0.1,1.28,-0.9)] = +0.608$

$\ell_3 = 0.2[g(0.1,1.28,-0.9)] = +0.216$

$k_4 = 0.2[f(0.2,1.608,-0.784)] = +0.846$

$\ell_4 = 0.2[g(0.2,1.608,-0.784)] = +0.242$

$\Delta x = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = +0.596$

$\Delta y = \frac{1}{6}(\ell_1 + 2\ell_2 + 2\ell_3 + \ell_4) = +0.212$

or

$$x(0.2) = +1.0 + 0.596 = 1.596 \qquad y(0.2) = -1.0 + 0.212 = -0.788$$

The comparable Taylor series (expanded about $t = 0$) for the given differential equations and initial conditions are

$$x(t) = +1.0 + 2t + 4t^2 + 4t^3 + \tfrac{10}{3}t^4 + \tfrac{33}{15}t^5 + \tfrac{109}{90}t^6 + \cdots \tag{10.140}$$

$$y(t) = -1.0 + t + \tfrac{4}{3}t^3 + t^4 + \tfrac{2}{3}t^5 + \tfrac{33}{90}t^6 + \cdots \tag{10.141}$$

For $t = 0.2$

$$x(0.2) = +1.5981145 \qquad y(0.2) = -0.7874968 \tag{10.142}$$

Table 10.9 summarizes the solutions for $t = 0.2$, 0.4, and 0.6. It is to be noted that all the values shown for the Taylor series are obtained by x, presuming $t = 0$ as the point of expansion. This was not the case for the Runge-Kutta solutions. There h was selected equal to 0.2, and the reference location was different for the three cases tabulated:

$$y(0.2) = y(0 + h = 0.2), \qquad y(0.4) = y(0.2 + h = 0.4)$$

and

$$y(0.6) = y(0.4 + h = 0.6)$$

Table 10.9 $\dfrac{dx}{dt} = t + 3x + y \qquad x(0) = +1.0$

$\dfrac{dy}{dt} = -2t + x \qquad y(0) = -1.0$

Function values	Taylor series†	Runge-Kutta‡
$x(0.2)$	+1.5981145	+1.5956
$y(0.2)$	−0.7874968	−0.7877
$x(0.4)$	+2.8088216	+2.77361
$y(0.4)$	−0.4807381	−0.45857
$x(0.6)$	+5.16358	+5.00435
$y(0.6)$	+0.0865472	+0.06974

† All Taylor series values are obtained by assuming expansion about $t = 0$.

‡ All Runge-Kutta solutions are obtained by assuming $h = 0.2$.

10.7.3 Higher-order Equations

It is demonstrated in Sec. 10.5 that higher-order differential equations can be treated as if they were sets of first-order equations. Runge-Kutta type forward integration solutions can be obtained, therefore, in the manner just described. It is evident, however, that such procedures are not the only ones available. More direct solutions can be obtained by repeating the whole process hereinbefore described for first-order cases.

The necessary Runge-Kutta equations for solution of a general second-order equation are summarized in Table 10.10. These are obtained by direct development, proceeding in the fashion described for the first-order case in Sec. 10.7.1.

Table 10.10 $y'' = f(x,y,y')$

$$k_1 = \frac{h^2}{2}\,[f(x,y,y')]$$

$$k_2 = \frac{h^2}{2}\left[f\left(x + \frac{1}{2}h,\ y + \frac{h}{2}y' + \frac{1}{4}k_1,\ y' + \frac{1}{h}k_1\right)\right]$$

$$k_3 = \frac{h^2}{2}\left[f\left(x + \frac{1}{2}h,\ y + \frac{h}{2}y' + \frac{1}{4}k_1,\ y' + \frac{1}{h}k_2\right)\right]$$

$$k_4 = \frac{h^2}{2}\left[f\left(x + h,\ y + hy' + k_3,\ y' + \frac{2}{h}k_3\right)\right]$$

$$\Delta y = \frac{1}{3}(k_1 + k_2 + k_3)$$

$$\Delta y' = \frac{1}{3h}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$y(x + h) = y(x) + h[y'(x)] + \Delta y$$
$$y'(x + h) = y'(x) + \Delta y'$$

To illustrate the procedure of solution using the equations contained in Table 10.10, consider the solution of Eq. (10.143),

$$y'' = 3xyy' - (1 + x)y \tag{10.143}$$

where

$$y(0) = 0 \qquad \text{and} \qquad y'(0) = +1.0 \tag{10.144}$$

For the case where $h = 0.2$, the Runge-Kutta solution is obtained as follows:

$$k_1 = \frac{(0.2)^2}{2}[f(0,0,1.0)] = 0$$

$$k_2 = 0.02[f(0.1,0.1,1.0)] = -0.0016$$

$$k_3 = 0.02[f(0.1,0.1,0.992)] = -0.0016048$$

$$k_4 = 0.02[f(0.2,0.198,0.984)] = -0.0024142$$

$$\Delta y = \frac{1}{3}(0 - 0.0016 - 0.0016048) = -0.0010683$$

$$\Delta y' = \frac{1}{0.6}(0 - 0.0032 - 0.0032096 - 0.0024142) = -0.014706$$

$$\begin{aligned} y(0.2) &= y(0) + 0.2[y'(0)] + \Delta y \\ &= 0 + 0.2(+1.0) - 0.0010683 = +0.198932 \end{aligned} \tag{10.145}$$

$$\begin{aligned} y'(0.2) &= y'(0) + \Delta y' \\ &= +1.0 - 0.014706 = +0.985294 \end{aligned} \tag{10.146}$$

For comparison, a Taylor series solution of Eq. (10.143) proceeds in the following fashion. Let

$$z = y' \tag{10.147}$$

Then

$$z' = 3xyy' - (1 + x)y \tag{10.148}$$

Independent and dependent boundary conditions are developed in Table 10.11.

Table 10.11

$y'(x) = z$	$z'(x) = 3xyy' - (1 + x)y$
$y(0) = 0$ [boundary condition]	$z(0) = +1.0$ [boundary condition]
$y'(0) = z(0) = +1.0$	$z'(0) = 3xyy' - (1 + x)y = 0$
$y''(0) = z'(0) = 0$	$z''(0) = -1.0$
$y'''(0) = z''(0) = -1.0$	$z'''(0) = +4.0$
$y''''(0) = z'''(0) = +4.0$	$z''''(0) = -1.0$
. . . .	. . . .

Since

$$y = y(0) + y'(0)x + \frac{y''(0)}{2!}x^2 + \frac{y'''(0)}{3!}x^3 + \frac{y''''(0)}{4!}x^4 + \cdots$$

and

$$z = z(0) + z'(0)x + \frac{z''(0)}{2!}x^2 + \frac{z'''(0)}{3!}x^3 + \frac{z''''(0)}{4!}x^4 + \cdots$$

the series representations are

$$y = x - \tfrac{1}{6}x^3 + \tfrac{1}{6}x^4 - \tfrac{1}{120}x^5 + \cdots \tag{10.149}$$

and

$$y' = z = +1.0 - \tfrac{1}{2}x^2 + \tfrac{2}{3}x^3 - \tfrac{1}{24}x^4 + \cdots \tag{10.150}$$

This gives for comparison with Eqs. (10.145) and (10.146) the following series-determined values at $x = 0.2$:

$$y(0.2) = +0.19893 \tag{10.151}$$

and

$$y'(0.2) = +0.98527 \tag{10.152}$$

The comparisons for these and other values of h are shown in Table 10.12.

Table 10.12

x function values	y and y' function values	Taylor series solution using Eqs. (10.149) and (10.150)	Solution using Runge-Kutta equations, Table 10.10
$x = 0.2$	$y(0.2) =$ $y'(0.2) =$	+0.19893 +0.98527	+0.19893 +0.98529
$x = 0.4$	$y(0.4) =$ $y'(0.4) =$	+0.39352 +0.96163	+0.39352 +0.96346
$x = 0.6$	$y(0.6) =$ $y'(0.6) =$	+0.58496 +0.96860	+0.58502 +0.95351

10.8 Analytical Continuation

It is evident from the preceding examples in this chapter that accuracy of solution can be obtained by the methods heretofore developed only if the range of independent variable under study is kept small. This is due primarily to the fact that as the size of the increment is increased, more terms must be included to maintain the same degree of accuracy. This almost always involves a vastly increased amount of computation. To circumvent this situation, initial-value problems are normally solved as two-stage problems: first, starting the solution, and second, analytical continuation.

The starting procedure has as its primary purpose the establishment of a few accurate values of $y = f(x)$ at preselected pivotal values of x. Since when x is held to a small interval, "accuracy" can be obtained with relatively few terms included, almost any of the methods hereinbefore described can be used for starting the solution.

The method selected for continuing the solution must be such that the accuracy of the first few values (obtained as described in the preceding paragraph) is maintained. Moreover, the method should not be too awkward to use, and must allow solution over an extended range of x. Procedures of this general type are illustrated in Fig. 10.3. They are frequently referred to as *marching methods*, since each step overlaps the others in such a way that only one value of y is unknown at a given time.

In general, methods for analytical continuation are of the finite-difference type; that is, they involve the use of function values (or their corresponding differences) at pivotal points. Two distinct and separate procedures have been extensively used: (1) predictor-corrector methods and (2) collocation methods.

10.8.1 Predictor-Corrector Methods

The basic concept of a predictor-corrector solution requires a two-stage operation. The first stage is an approximate determination of the solution using a so-called "open integration formula," that is, an integration formula that involves known values "to the left" of the point being sought. The second stage makes use of the result of the first calculations in a second, more exact, "closed integration formula."

To illustrate the basic concepts involved, consider the first-order differential equation

$$\frac{dy}{dx} = f(x,y) \tag{10.153}$$

It is assumed that the range of solution of interest is from x_0 to $x_0 + h = x_1$. It is also presumed that the Euler and Euler-Cauchy equations from Table 10.5 are available. The first of these,

$$y(x_1) = y(x_0) + \Delta y = y(x_0) + h[f(x_0,y_0)] \tag{10.154}$$

is of the "open" type. (Note that all values on the right-hand side of the equation are concerned with location x_0, not x_1.) The Euler-Cauchy equation written in the same form is

$$y(x_1) = y(x_0) + \tfrac{1}{2}\{h[f(x_0,y_0)] + h[f(x_1,y(x_1))]\} \tag{10.155}$$

It is to be observed that in this case both sides of Eq. (10.155) involve $y(x_1)$. It is therefore said to be of the "closed" type. Referring to Table 10.5 shows that the error associated with Eq. (10.154) is of the order of h^2, whereas that for Eq. (10.155) is of the order of h^3.

It is possible to define a well-ordered iterative-solution procedure by using Eqs. (10.154) and (10.155). If the predictor value is denoted with an asterisk, "*", the two equations are:

Predictor: $$y(x_1)^* = y(x_0) + h[f(x_0,y_0)] \tag{10.156}$$

Corrector: $$y(x_1) = y(x_0) + \tfrac{1}{2}h[f(x_0,y_0) + f(x_1,y(x_1)^*)] \tag{10.157}$$

First, $y(x_1)^*$ would be predicted from (10.156). This value would then be substituted into Eq. (10.157) and a new $y(x_1)$ would be computed. Should the predicted and computed values differ too greatly, Eq. (10.157) would again be used, but this time with the $y(x_1)$ just computed as a new predicted value. This iteration process would be continued until the desired accuracy was achieved.

The best known of the predictor-corrector methods are those due to Milne. In general, they are based on integration formulas of the type considered in Chap. 9, which are summarized in Tables 9.29 and 9.30.

By considering only those integration formulas in Table 9.29 whose errors are of the order $h^5f''''(\xi)$, the following can be listed:

OPEN EQUATIONS

From Eq. (10), Table 9.29:

$$y_{+1} = y_{-3} + 4h\frac{2f_{-2} - f_{-1} + 2f_0}{3} + \frac{14}{45}h^5f''''(\xi) \tag{10.158}$$

From Eq. (13), Table 9.29:

$$y_{+1} = y_{-4} + 5h\frac{11f_{-3} + f_{-2} + f_{-1} + 11f_0}{24} + \frac{95}{144}h^5f''''(\xi) \tag{10.159}$$

CLOSED EQUATIONS

From Eq. (4), Table 9.29:

$$y_{+1} = y_{-1} + 2h\frac{f_{-1} + 4f_0 + f_{+1}}{6} - \frac{1}{90}h^5f''''(\xi) \tag{10.160}$$

From Eq. (8), Table 9.29:

$$y_{+1} = y_{-2} + 3h\frac{f_{-2} + 3f_{-1} + 3f_0 + f_{+1}}{8} + \frac{3}{80}h^5f''''(\xi) \tag{10.161}$$

Comparing the error terms shows that Eq. (10.158) is clearly superior to Eq. (10.159) as a predictor equation. Similarly, Eq. (10.160) should be selected as the corrector formula. [Note that the error in the selected corrector equation (10.160) is less than that in the predictor equation

(10.158).] Therefore,

$$y_{+1} = y_{-3} + 4h\,\frac{2f_{-2} - f_{-1} + 2f_0}{3} \qquad \text{(Predictor)} \qquad (10.162)$$

$$y_{+1} = y_{-1} + h\,\frac{f_{-1} + 4f_0 + f_{+1}}{3} \qquad \text{(Corrector)} \qquad (10.163)$$

Predictor-corrector methods and formulas are also defined in terms of *differences*. The Adams-Bashford method, for example, expresses the equations in terms of backward differences.

$$y_{+1} = y_0 + h(f_0 + \tfrac{1}{2}\nabla f_0 + \tfrac{5}{12}\nabla^2 f_0 + \tfrac{3}{8}\nabla^3 f_0 + \tfrac{251}{720}\nabla^4 f_0 + \tfrac{95}{288}\nabla^5 f_0 + \cdots) \qquad \text{(Predictor)} \qquad (10.164)$$

$$y_{+1} = y_0 + h(f_1 - \tfrac{1}{2}\nabla f_1 - \tfrac{1}{12}\nabla^2 f_1 - \tfrac{1}{24}\nabla^3 f_1 - \tfrac{19}{720}\nabla^4 f_1 - \tfrac{3}{160}\nabla^5 f_1 - \cdots) \qquad \text{(Corrector)} \qquad (10.165)$$

As an illustration of the use of Milne's method, consider the continuation of the solution to

$$\frac{dy}{dx} = 1 - xy = f(x,y) \qquad (10.166)$$

where the three initial points were obtained by Heun's method and are given in Table 10.13. Only the definition of y at $x = 0.8$ is illustrated.

Table 10.13

x	y	$f(x,y) = \dfrac{dy}{dx}$	
0	1.0000	1.0000	← (Given boundary condition)
0.2	1.1775	0.7645	(Obtained by using Heun's method)
0.4	1.3020	0.4792	
0.6	1.3672	0.1797	

$$y(0.8) = y(0) + 4h\,\frac{2f(0.2,1.1775) - f(0.4,1.3020) + 2f(0.6,1.3672)}{3}$$

$$= 1.0000 + \frac{4}{3}(0.2)[2(0.7654) - (0.4792) + 2(0.1797)]$$

$$= 1.0000 + 0.3758 = 1.3758 \qquad \text{(Predictor)} \qquad (10.167)$$

Substituting this value into the corrector equation,

$$y(0.8) = y(0.4) + h\,\frac{f(0.4,1.3020) + 4f(0.6,1.3672) + f(0.8,1.3758)}{3}$$

$$= 1.3020 + \tfrac{1}{3}(0.2)[(0.4792) + 4(0.1797) - (0.1006)]$$

$$= 1.3020 + 0.0667(1.0978) = 1.3752 \qquad \text{(Corrector)} \qquad (10.168)$$

Comparison would now be made between the values obtained from the corrector equation (10.168) and the predicted value from Eq. (10.167). Should these be sufficiently close, $y(0.8) = 1.3752$ would be retained as the value to use for further extensions of the range of solution. Had there been too great a difference between the two values, a second application of the corrector equation (10.163) would have been carried out. This time, however, $y(0.8) = 1.3752$ would have been used as the predictor value.

It is evident that Milne's predictor-corrector method can be extended directly to the solution of sets of first-order differential equations. Assume, for example, that the two equations listed as (10.169) are prescribed.

$$\frac{dx}{dt} = f(x,y,t) \qquad \frac{dy}{dt} = g(x,y,t) \qquad (10.169)$$

Further, presume that four starting values, at equally spaced pivotal points, have been established by some other method.

$$\left|\begin{array}{l|c|c||c|c|}
t_0 & x_0 & y_0 & f_0 & g_0 \\
t_1(= t_0 + h) & x_1 & y_1 & f_1 & g_1 \\
t_2(= t_1 + h) & x_2 & y_2 & f_2 & g_2 \\
t_3(= t_2 + h) & x_3 & y_3 & f_3 & g_3
\end{array}\right.$$

The predictor values are

$$x_4 = x_0 + \tfrac{4}{3}h(2f_1 - f_2 + 2f_3) \qquad \text{(Predictor)} \qquad (10.170)$$

$$y_4 = y_0 + \tfrac{4}{3}h(2g_1 - g_2 + 2g_3) \qquad \text{(Predictor)} \qquad (10.171)$$

Correspondingly, the corrector equations are

$$x_4 = x_2 + \tfrac{1}{3}h(f_2 + 4f_3 + f_4) \qquad \text{(Corrector)} \qquad (10.172)$$

$$y_4 = y_2 + \tfrac{1}{3}h(g_2 + 4g_3 + g_4) \qquad \text{(Corrector)} \qquad (10.173)$$

Milne's method can also be extended to facilitate solution of equations of higher order. Consider, for example, the general second-order equation

$$y'' = f(x,y,y') \qquad (10.174)$$

As in the preceding cases, it is assumed that sufficiently accurate values of y, y', and y'' are available for at least four consecutive, equally spaced pivotal locations. If these locations are denoted by the subscripts -3, -2, -1, and 0, the first derivative of y with respect to x at location $+1$ can be predicted from the open-type equation

$$y'^{*}_{+1} = y'_{-3} + \tfrac{4}{3}h(2y''_{-2} - y''_{-1} + 2y''_{0}) \tag{10.175}$$

With this, a value for y_{+1} can also be predicted.

$$y^{*}_{+1} = y_{-1} + \tfrac{1}{3}h(y'_{-1} + 4y'_{0} + y'^{*}_{+1}) \tag{10.176}$$

A predicted value for the second derivative is

$$y''^{*}_{+1} = f(x_1, y^{*}_1, y'^{*}_1) \tag{10.177}$$

A more nearly correct solution for y'_{+1} (and, in turn, y_{+1}) can be obtained by using the closed formula

$$y'_{+1} = y'_{-1} + \tfrac{1}{3}h(y''_{-1} + 4y''_{0} + y''^{*}_{+1}) \tag{10.178}$$

The process would continue using Eqs. (10.176) to (10.178) until the desired accuracy had been achieved.

10.8.2 Collocation Method

The collocation method requires that the differential equation be satisfied to the degree of accuracy desired only at the pivotal points. It does not specify what is the situation between these pivotal locations, nor does it give an indication of the error in the integral. Nevertheless, consider the solution of the same first-order equation that was solved in the preceding section, using the difference expressions contained in Table 9.28. The given differential equation is

$$\frac{dy}{dx} = 1 - xy \tag{10.179}$$

If a fourth-order parabola is assumed, the first derivative can be expressed in terms of five pivotal values (four known, and one being sought). From Table 9.28,

$$D(y_0) = \frac{1}{24h}(6y_{-4} - 32y_{-3} + 72y_{-2} - 96y_{-1} + 50y_0) + \tfrac{1}{5}h^4y_0^{\mathrm{V}} \tag{10.180}$$

By selecting location 0 as the location of the desired "new solution," the differential equation is written in terms of finite differences as shown in Eq. (10.181).

$$\frac{1}{(24)(0.2)}[6y(0) - 32y(0.2) + 72y(0.4) - 96y(0.6) + 50y(0.8)] = 1 - 0.8[y(0.8)]$$

or

$$\frac{1}{(24)(0.2)}[6(1.0000) - 32(1.1775) + 72(1.3020) - 96(1.3672) + 50y(0.8)] = 1 - 0.8[y(0.8)] \qquad (10.181)$$

This gives as an answer

$$y(0.8) = 1.3745 \qquad (10.182)$$

As a second example using the collocation method, consider the second-order differential equation

$$\frac{d^2y}{dx^2} + x\left(\frac{dy}{dx}\right) + y + C = 0 \qquad (10.183)$$

If a third-order system is assumed, the necessary backward-difference equations from Table 9.28 are

$$\left.\frac{d^2y}{dx^2}\right|_{x_0} = \frac{1}{h^2}(-y_{-3} + 4y_{-2} - 5y_{-1} + 2y_0) \qquad (10.184)$$

$$\left.\frac{dy}{dx}\right|_{x_0} = \frac{1}{6h}(-2y_{-3} + 9y_{-2} - 18y_{-1} + 11y_0) \qquad (10.185)$$

$$\left.y\right|_{x_0} = y_0 \qquad (10.186)$$

Expansion about the unknown point x_0 has been assumed.

Substituting the finite-difference expressions of Eqs. (10.184) and (10.185) into the differential equation (10.183) yields

$$\frac{1}{h^2}(-y_{-3} + 4y_{-2} - 5y_{-1} + 2y_0) + x_0\left[\frac{1}{6h}(-2y_{-3} + 9y_{-2} - 18y_{-1} + 11y_0)\right] + y_0 + C = 0 \qquad (10.187)$$

or

$$y_0(2 + \tfrac{11}{6}x_0h + h^2) = y_{-3}(1 + \tfrac{1}{3}x_0h) - y_{-2}(4 + \tfrac{3}{2}x_0h) + y_{-1}(5 + 3x_0h) - Ch^2 \qquad (10.188)$$

10.9 Stability of Numerical Solutions to Initial-value Problems

Because of the general nature of this introductory text, only a minimum amount of information concerning errors has been included. From that given, it might be assumed that the error associated with a particular calculated pivotal value is dependent only on the "local truncation error" of that calculational step. This is not necessarily true. In the solution of initial-value problems, particularly, round-off and truncation effects of previous calculations can markedly influence future errors, and when these build up rapidly, instability results. It should be noted that this type of numerical instability can occur even when the differential equation in question is inherently stable.

Certain systems of equations and boundary conditions are inherently unstable, a given initial condition only partially (because of errors) suppressing an undesired major mode of solution. As the independent variable increases, the influence of these initial errors can completely dominate the solution. These "ill-conditioned" systems are characterized by the condition that a small change in the initial condition will result in a large change in the solution.

For certain methods of solution, notably, the truncated Taylor series method and the Runge-Kutta method, instability occurs when h exceeds a certain limiting value. The particular limiting value depends on the equation being examined. However, no instability need be feared if h is kept "sufficiently small."

A third condition leading possibly to instability results from the application of a difference equation of higher order than that of the differential equation being examined. These higher-order difference equations may introduce "spurious solutions" that have nothing to do with the equation in question, and yet they may completely dominate the solution. Predictor-corrector methods may result in this type of instability. It is to be noted, however, that here stability depends on how the process is used. If the iteration converges, and the process is continued until the corrector is "exactly satisfied," the numerical process is stable.

10.10 Boundary-value Problems

As was pointed out earlier in this chapter, boundary-value problems are those for which constraints or conditions on the solution are specified at more than one value of the independent variable. Normally, these occur at "end points," but this need not be the case.

Conceptually, all the methods that were hereinbefore developed for the solution of initial-value problems can be used to solve boundary-value problems. It should be noted, however, that such direct extension to this class of problems normally requires a trial-and-error procedure starting from certain presumptions regarding the undefined initial conditions.

The collocation method, which requires that the differential equation in difference form be satisfied only at the pivotal locations, is ideally suited to the solution of linear boundary-value problems. Consider, for exam-

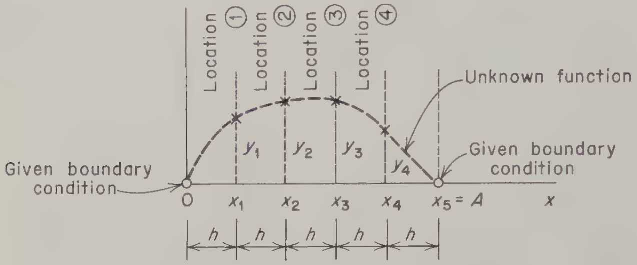


Figure 10.9

ple, the solution to the differential equation

$$\frac{d^2y}{dx^2} + f_1(x) = 0 \tag{10.189}$$

The range of interest of the independent variable is $0 \leq x \leq A$. The given boundary conditions are

$$\begin{aligned} &\text{At } x = 0: \qquad && y = 0 \\ &\text{At } x = A: \qquad && y = 0 \end{aligned} \tag{10.190}$$

If it is assumed that a sufficiently accurate solution can be obtained by subdividing the range in question into five equal parts, that is, $h = \frac{1}{5}A$, the problem can be described as shown in Fig. 10.9. The desired function values at x_1, x_2, x_3, and x_4 are, respectively, y_1, y_2, y_3, and y_4.

From Table 9.27, by assuming a second-order system with symmetrical placement of pivotal points on either side of the location in question, i, the second derivative of the function can be approximated by the expression

$$D^2(y_i) = \frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) \tag{10.191}$$

In difference form, then, the given differential equation (10.189) is

$$\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) + f_1(x_i) = 0$$

or

$$(y_{i-1} - 2y_i + y_{i+1}) + h^2 f_1(x_i) = 0 \tag{10.192}$$

At location ①—where $x_i = x_1$, $y_{i-1} = y_0$, $y_i = y_1$, and $y_{i+1} = y_2$—the necessary condition is

$$y_0 - 2y_1 + y_2 + h^2 f_1(x_1) = 0 \tag{10.193}$$

But the given boundary condition $y(0) = 0$ requires that $y_0 = 0$. Therefore Eq. (10.193) can be written as

$$-2y_1 + y_2 + h^2 f_1(x_1) = 0 \tag{10.194}$$

Expansion about each of the other pivotal locations will result in a similar linear algebraic equation. The total set is given by

$$\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -h^2 f_1(x_1) \\ -h^2 f_1(x_2) \\ -h^2 f_1(x_3) \\ -h^2 f_1(x_4) \end{bmatrix} \tag{10.195}$$

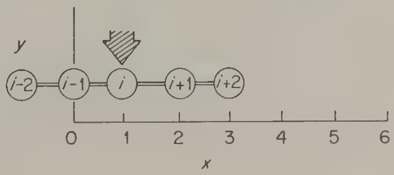
Solution can be achieved by any of the methods defined in Chap. 5.

The reduction of the original differential equation to a set of n linear algebraic equations is characteristic of solutions of linear boundary-value problems using the collocation method. The unknowns are the function values at the pivotal locations.

Had the differential equation in question been of the fourth order, it would have been necessary to express the fourth derivative in terms of five adjacent pivotal locations. Expansion about location ① leads to the condition illustrated in Fig. 10.10. Note that location $i - 2$ falls completely outside the range of variable of interest. More than this, had it been necessary to ensure that the differential equation (in difference form) was satisfied at location 0, both $i - 2$ and $i - 1$ would have been "off." One method for handling such situations is illustrated in Fig. 10.11. Expansion about location 0 presumes a difference equation written in terms of pivotal values at 0 and forward of it. For the second case, expansion about location 1, it has been presumed that equations are available which allow definition of derivatives about a location one removed from the leftmost pivotal point—so-called "off-center" forward-difference expressions. The third, fourth, and fifth cases illustrated would be written in terms of pivotal values symmetrically placed with respect to the reference locations.

A second—more often used—method for handling problems of this type is to first write all equations in symmetrical form, presuming a fictitious

Figure 10.10



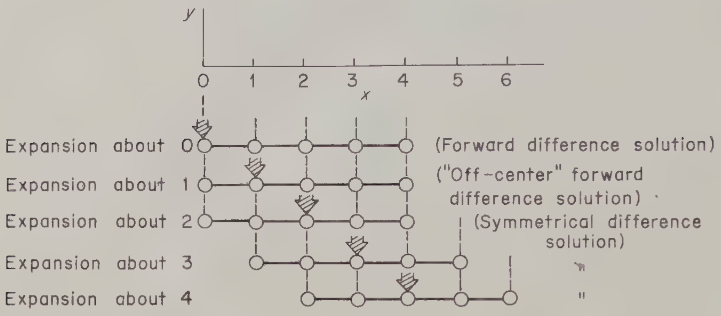

y
0
1
2
3
4
5
6
x
Expansion about 0
Expansion about 1
Expansion about 2
Expansion about 3
Expansion about 4
(Forward difference solution)
("Off-center" forward difference solution)
(Symmetrical difference solution)
"
"

Figure 10.11

range of independent variable with corresponding unknown pivotal values at fictitious locations. The relationships of the pivotal values in the fictitious range to those in the range of interest would be defined by the specified boundary conditions of the problem in question.

To illustrate how such a procedure works, assume that a fourth-order system is specified, and that boundary conditions at 0 are defined as follows:

$$y_0 = q \tag{10.196}$$

$$D(y_0) = 0 \tag{10.197}$$

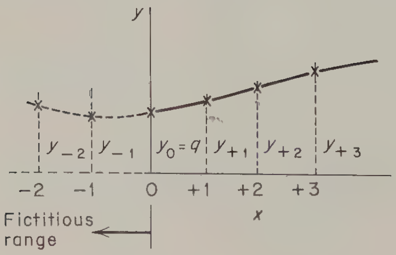
$$D^3(y_0) = 0 \tag{10.198}$$

The pivotal values in question are shown in Fig. 10.12, as is the assumed fictitious range. From Table 9.27,

$$D(y_0) = \frac{1}{12h}(y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2}) \tag{10.199}$$

$$D^3(y_0) = \frac{1}{2h^3}(-y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2}) \tag{10.200}$$

Figure 10.12



Substituting these into the specified boundary conditions gives three simultaneous equations in the unknowns y_{-2}, y_{-1}, and y_0.

$$\begin{aligned} y_0 &= q \\ y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2} &= 0 \\ -y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2} &= 0 \end{aligned} \qquad (10.201)$$

Solution of this set of equations requires that $y_{-2} = y_{+2}$, $y_{-1} = y_{+1}$, and $y_0 = q$.

It is to be noted that boundary conditions are not always related to only one of the derivatives. For example, the fourth boundary condition specified in Example 7.4 is

$$y''' + k^2y' = 0$$

or

$$D^3(y_0) + k^2\,D(y_0) = 0 \qquad (10.202)$$

For this case the corresponding difference expression is

$$\frac{1}{2h^3}(-y_{-2} + 2y_{-1} - 2y_{+1} + y_{+2}) + \frac{k^2}{12h}(y_{-2} - 8y_{-1} + 8y_{+1} - y_{+2}) = 0 \qquad (10.203)$$

or

$$y_{-2}(1 - \alpha) - y_{-1}(2 - 8\alpha) + y_{+1}(2 - 8\alpha) - y_{+2}(1 - \alpha) = 0 \qquad (10.204)$$

where

$$\alpha = \frac{k^2h^2}{6} \qquad (10.205)$$

The two most often encountered pairs of boundary conditions are

$$y = 0 \qquad \frac{dy}{dx} = 0 \qquad (10.206)$$

and

$$y = 0 \qquad \frac{d^2y}{dx^2} = 0 \qquad (10.207)$$

The first of these is illustrated in Fig. 10.13. It is to be noted that solu-

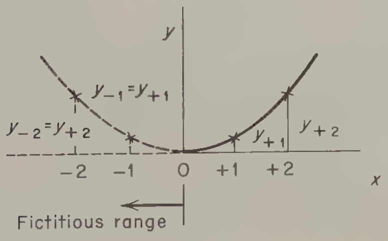


Figure 10.13

Boundary conditions

$y_0 = 0$

$D(y_0) = 0$

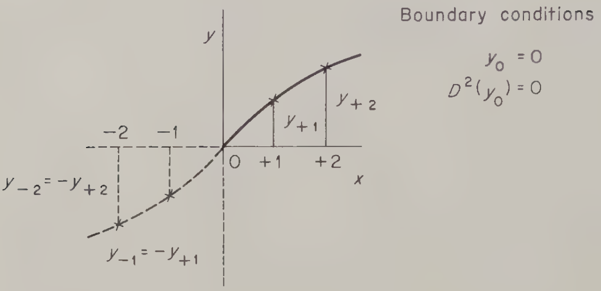


Figure 10.14

tion requires the following relationships:

$$y_0 = 0 \qquad y_{-1} = y_{+1} \qquad y_{-2} = y_{+2} \tag{10.208}$$

For the second set of boundary conditions—$y = 0$ and $d^2y/dx^2 = 0$—the relationships illustrated in Fig. 10.14 would be required.

To illustrate the application of these procedures, consider the solution to the following fourth-order differential equation:

$$\frac{d^4y}{dx^4} + C_1 \frac{d^2y}{dx^2} + C_2 = 0 \tag{10.209}$$

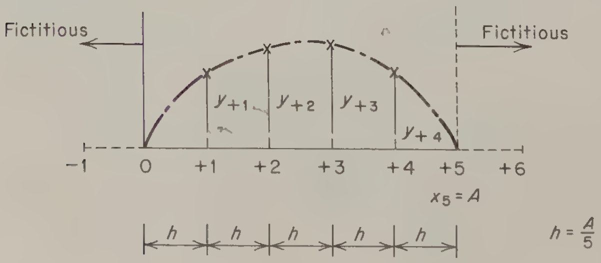
where the specified boundary conditions are

$$\text{At } x = 0: \qquad y = 0 \quad \text{and} \quad \frac{dy}{dx} = 0 \tag{10.210}$$

$$\text{At } x = A: \qquad y = 0 \quad \text{and} \quad \frac{d^2y}{dx^2} = 0 \tag{10.211}$$

It is to be assumed that the desired accuracy can be achieved by subdividing the range of x of interest into five equal parts, as shown in Fig. 10.15. For completeness, the two fictitious regions, one at each end of

Figure 10.15



the range of interest, have also been indicated. The difference equation corresponding to the given differential equation (10.209) is

$$\frac{1}{h^4}(y_{i-2} - 4y_{i-1} + 6y_i - 4y_{i+1} + y_{i+2}) + C_1\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) + C_2 = 0 \tag{10.212}$$

or

$$y_{i-2}(+1) + y_{i-1}(C_1h^2 - 4) + y_i(-2C_1h^2 + 6) + y_{i+1}(C_1h^2 - 4) + y_{i+2}(+1) = -C_2h^4 \tag{10.213}$$

It is to be noted that in Eq. (10.212) a fourth-order (five points) polynomial is used to express the fourth derivative, whereas a second-order (three points) polynomial is used for the second derivative. This was done to keep the order of the error terms the same (both are of the order of h^2).

The boundary conditions require that pivotal values at the fictitious locations be related to those in the range in question according to the following:

$$y_{-1} = y_{+1} \qquad \text{and} \qquad y_{+6} = -y_{+4} \tag{10.214}$$

The governing matrix is therefore

$$\begin{bmatrix} -2C_1h^2 + 7 & C_1h^2 - 4 & +1 & 0 \\ C_1h^2 - 4 & -2C_1h^2 + 6 & C_1h^2 - 4 & +1 \\ +1 & C_1h^2 - 4 & -2C_1h^2 + 6 & C_1h^2 - 4 \\ 0 & +1 & C_1h^2 - 4 & -2C_1h^2 + 5 \end{bmatrix} \begin{bmatrix} y_{+1} \\ y_{+2} \\ y_{+3} \\ y_{+4} \end{bmatrix} = \begin{bmatrix} -C_2h^4 \\ -C_2h^4 \\ -C_2h^4 \\ -C_2h^4 \end{bmatrix} \tag{10.215}$$

The solution procedure just demonstrated—where, by using the boundary conditions, fictitious pivotal values are first defined in terms of values within the range in question—is not the only procedure available. The problem could equally well have been formulated in terms of eight unknown pivotal values: y_{-1}, y_0, y_{+1}, y_{+2}, y_{+3}, y_{+4}, y_{+5}, and y_{+6}. The four additional equations would correspond to the specified boundary conditions.

$$
\begin{bmatrix}
-1 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \\
0 & +1 & 0 & 0 & 0 & 0 & 0 & 0 \\
+1 & C_1h^2-4 & -2C_1h^2+6 & C_1h^2-4 & +1 & 0 & 0 & 0 \\
0 & +1 & C_1h^2-4 & -2C_1h^2+6 & C_1h^2-4 & +1 & 0 & 0 \\
0 & 0 & +1 & C_1h^2-4 & -2C_1h^2+6 & C_1h^2-4 & +1 & 0 \\
0 & 0 & 0 & +1 & C_1h^2-4 & -2C_1h^2+6 & C_1h^2-4 & +1 \\
0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\
0 & 0 & 0 & 0 & 0 & +1 & 0 & +1
\end{bmatrix}
\begin{bmatrix}
y_{-1} \\ y_0 \\ y_{+1} \\ y_{+2} \\ y_{+3} \\ y_{+4} \\ y_{+5} \\ y_{+6}
\end{bmatrix}
=
\begin{bmatrix}
0 \\ 0 \\ -C_2h^2 \\ -C_2h^2 \\ -C_2h^2 \\ -C_2h^2 \\ 0 \\ 0
\end{bmatrix}
\tag{10.216}
$$

It is to be noted that this matrix is not symmetrical.

The collocation method can be equally well adapted to the solution of sets of boundary-value differential-equation problems. Consider, for example, the two equations

$$\frac{d^2x}{dt^2} + 7\frac{dy}{dt} - 3x = 13 \tag{10.217}$$

$$\frac{d^2y}{dt^2} - x = 0 \tag{10.218}$$

The given boundary conditions are

$$\begin{array}{llll} \text{At } t = 0: & y = 0 & \text{and} & x = 0 \\ \text{At } t = A: & y = 0 & \text{and} & x = b \end{array} \tag{10.219}$$

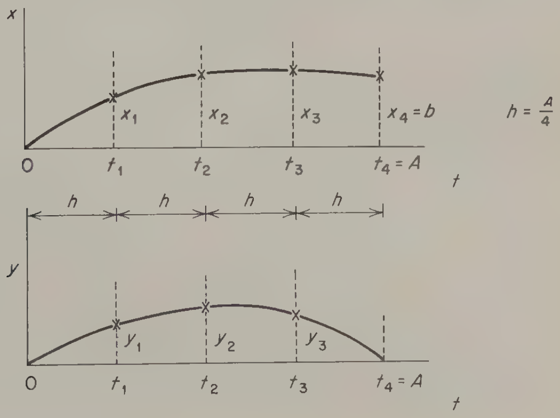
If it is assumed that a sufficiently accurate solution can be obtained by dividing the range of independent variable in question into four equal parts, the unknown pivotal values are those shown in Fig. 10.16. Writing the first of the two differential equations in symmetrical difference form, assuming a second-order approximating polynomial, yields the following difference equation:

$$\frac{1}{h^2}(x_{i-1} - 2x_i + x_{i+1}) + 7/2h(-y_{i-1} + y_{i+1}) - 3x_i = 13 \tag{10.220}$$

or

$$\begin{aligned}(x_{i-1})(+1) + x_i(-2 - 3h^2) + x_{i+1}(+1) \\ + y_{i-1}(-7/2h) + y_{i+1}(+7/2h) = +13h^2\end{aligned} \tag{10.221}$$

Figure 10.16



For Eq. (10.218), the corresponding difference equation is

$$x_i(-h^2) + y_{i-1}(+1) + y_i(-2) + y_{i+1}(+1) = 0 \qquad (10.222)$$

The equations controlling the solution are

$$\begin{bmatrix} -2 - 3h^2 & +1 & 0 & 0 & +7/2h & 0 \\ +1 & -2 - 3h^2 & +1 & -7/2h & 0 & +7/2h \\ 0 & +1 & -2 - 3h^2 & 0 & -7/2h & 0 \\ -h^2 & 0 & 0 & -2 & +1 & 0 \\ 0 & -h^2 & 0 & +1 & -2 & +1 \\ 0 & 0 & -h^2 & 0 & +1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} +13h^2 \\ +13h^2 \\ +13h^2 - b \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (10.223)$$

10.11 Eigenvalue Problems

Eigenvalue problems (or characteristic-value problems) can also be solved by using the collocation method just described. Consider, for example, the solution to Eqs. (10.224) and (10.225).

$$\frac{d^2y}{dx^2} + \lambda^2 y = 0 \qquad (10.224)$$

$$\begin{aligned} &\text{At } x = 0: && y = 0 \\ &\text{At } x = L: && y = 0 \end{aligned} \qquad (10.225)$$

By recalling that the symmetrical difference expression for the second derivative is

$$\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1})$$

the differential equation can be written in difference form as

$$\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) + \lambda^2 y_i = 0$$

or

$$y_{i-1} - 2y_i + y_{i+1} + (\lambda^2 h^2) y_i = 0$$

or

$$y_{i-1}(+1) + y_i(-2 + \lambda^2 h^2) + y_{i+1}(+1) = 0 \qquad (10.226)$$

If the range of independent variable in question (that is, L) is divided into two segments ($n = 2$), the problem corresponds to that illustrated in Fig. 10.17.

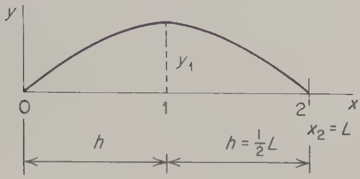


Figure 10.17

Writing Eq. (10.226) about location ① yields

$$y_0(+1) + y_1(-2 + \lambda^2 h^2) + y_2(+1) = 0$$

But $y_0 = y_2 = 0$. Therefore,

$$y_1(-2 + \lambda^2 h^2) = 0 \tag{10.227}$$

For a nontrivial solution, it is necessary that

$$-2 + \lambda^2 h^2 = 0$$

The eigenvalue solution for this one-degree-of-freedom assumption is, therefore,

$$\lambda^2_{n=2} = \frac{2}{h^2} = \frac{8}{L^2} \tag{10.228}$$

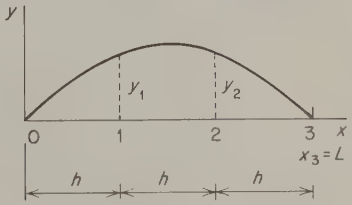
If the region is subdivided into three segments ($n = 3$) (see Fig. 10.18), solution proceeds in the following fashion:
Expansion about location ① gives

$$y_1(-2 + \lambda^2 h^2) + y_2(+1) = 0 \tag{10.229}$$

Expansion about location ② gives

$$y_1(+1) + y_2(-2 + \lambda^2 h^2) = 0 \tag{10.230}$$

Figure 10.18



In matrix form, the controlling set of equations is

$$\begin{bmatrix} -2+\lambda^2h^2 & +1 \\ +1 & -2+\lambda^2h^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{10.231}$$

Using the methods of Chap. 7 gives the smallest eigenvalue as

$$\lambda_{n=3}^2 = \frac{9}{L^2} \tag{10.232}$$

The governing set of equations for the case where $n = 4$ (that is, $h = L/4$) is

$$\begin{bmatrix} -2+\lambda^2h^2 & +1 & 0 \\ +1 & -2+\lambda^2h^2 & +1 \\ 0 & +1 & -2+\lambda^2h^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = [0] \tag{10.233}$$

Here, the smallest eigenvalue is found to be

$$\lambda_{n=4}^2 = \frac{9.36}{L^2} \tag{10.234}$$

and the corresponding eigenvector is

$$\begin{bmatrix} +0.707 \\ +1.000 \\ +0.707 \end{bmatrix} \tag{10.235}$$

For an indication of the error associated with each of these solutions, consider the nondimensional plot of Fig. 10.19. Note that both of the coordinate dimensions are "squared."

As a second example, consider the solution of the more involved eigenvalue problem defined by Eqs. (10.236) to (10.238).

$$\frac{d^4y}{dx^4} + k^2\frac{d^2y}{dx^2} = 0 \tag{10.236}$$

At $x = 0$: $\quad y = 0 \quad$ and $\quad y' = 0 \qquad$ (10.237)

At $x = L$: $\quad y'' = 0 \quad$ and $\quad y''' + k^2y' = 0 \qquad$ (10.238)

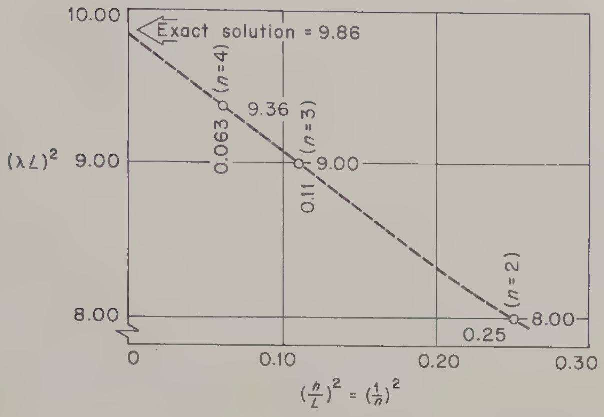


Figure 10.19

It is assumed that sufficient accuracy can be realized by selecting $n = 4$. The problem is defined graphically in Fig. 10.20. Two fictitious ranges have been presumed: one to the left of location ⓪, the other to the right of location ④.

From Table 9.27, the following difference expressions for derivatives are obtained:

$$D^3(y_i) = \frac{1}{2h^3}(-y_{i-2} + 2y_{i-1} - 2y_{i+1} + y_{i+2}) \tag{10.239}$$

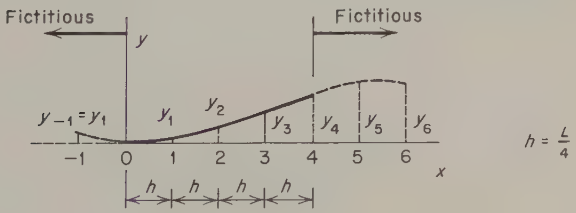
$$D^2(y_i) = \frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) \tag{10.240}$$

$$D(y_i) = \frac{1}{2h}(-y_{i-1} + y_{i+1}) \tag{10.241}$$

To establish the relationships in the fictitious region to the left, consider boundary conditions (10.237): $y(0) = 0$ and $y'(0) = 0$. These give

$$y_0 = 0 \qquad y_{-1} = y_{+1} \tag{10.242}$$

Figure 10.20



For y_5 and y_6, it is necessary to consider the constraints $y''(L) = 0$, and $y'''(L) + k^2y'(L) = 0$. Using the first of these,

$$\frac{1}{h^2}(y_3 - 2y_4 + y_5) = 0$$

or

$$y_5 = 2y_4 - y_3 \tag{10.243}$$

For the second boundary condition, $y'''(L) + k^2y'(L) = 0$, the relationship is

$$\frac{1}{2h^3}(-y_2 + 2y_3 - 2y_5 + y_6) + \frac{k^2}{2h}(-y_3 + y_5) = 0$$

or

$$y_6 = y_2(+1) + y_3(-4 + 2\alpha) + y_4(4 - 2\alpha) \tag{10.244}$$

where

$$\alpha = h^2k^2$$

The difference equation corresponding to the given fourth-order differential equation is developed as follows:

$$\frac{d^4y}{dx^4} + k^2\frac{d^2y}{dx^2} = 0$$

$$\frac{1}{h^4}(y_{i-2} - 4y_{i-1} + 6y_i - 4y_{i+1} + y_{i+2}) + \frac{k^2}{h^2}(y_{i-1} - 2y_i + y_{i+1}) = 0 \tag{10.245}$$

or

$$y_{i-2}(+1) + y_{i-1}(-4 + \alpha) + y_i(6 - 2\alpha) + y_{i+1}(-4 + \alpha) + y_{i+2}(+1) = 0 \tag{10.246}$$

Writing Eq. (10.246) at each of the pivotal locations 1, 2, 3, and 4, taking into account the previously defined relationships for the fictitious pivotal points, yields the desired four simultaneous linear algebraic homogeneous equations.

$$\begin{bmatrix} 7 - 2\alpha & -4 + \alpha & +1 & 0 \\ -4 + \alpha & 6 - 2\alpha & -4 + \alpha & +1 \\ +1 & -4 + \alpha & 5 - 2\alpha & -2 + \alpha \\ 0 & +2 & -4 + \alpha & 2 - 2\alpha \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = [0] \tag{10.247}$$

The smallest eigenvalue associated with this 4 × 4 set is

$$\alpha_{n=4} = 0.0745 \tag{10.248}$$

Since $\alpha = h^2k^2$, this solution can equally well be expressed in terms of kL.

$$\alpha = h^2k^2 = \left(\frac{L}{n}\right)^2 k^2 \qquad \text{or} \qquad (kL)^2 = \alpha n^2 \tag{10.249}$$

where n is the number of equal-length segments into which the interval L is subdivided. Therefore,

$$(kL)^2_{n=4} = (0.0745)(4)^2 = 1.192 \tag{10.250}$$

This should be compared with the exact solution

$$(kL)^2_{\text{exact}} = 2.475 \tag{10.251}$$

It could be speculated that the large error associated with solution (10.250) is due to the dominating influence of the boundary conditions in the selected four-variable system. If eight segments are chosen, the first eigenvalue is

$$(kL)^2_{n=8} = (0.0384)(8)^2 = 2.460 \tag{10.252}$$

Table 10.14 summarizes the first three eigenvalues obtained in the manner described above, assuming both $n = 4$ and $n = 8$. For comparison, exact values have also been listed.

Table 10.14

n	First eigenvalue	Second eigenvalue	Third eigenvalue
4	$(kL)^2 = 1.19$	$(kL)^2 = 16.7$	$(kL)^2 = 44.2$
8	$(kL)^2 = 2.46$	$(kL)^2 = 21.6$	$(kL)^2 = 57.0$
Exact value	$(kL)^2 = 2.48$	$(kL)^2 = 22.3$	$(kL)^2 = 61.9$

10.12 Nonlinear Boundary-value Problems

It is evident from the preceding two sections of this chapter that the collocation method is ideally suited to the solution of linear boundary-value problems. Had the differential equations in question been nonlinear, there would have resulted sets of nonlinear algebraic equations.

For single nonlinear boundary-value problems, a trial-and-error initial-value approach, using the methods described earlier in this chapter, is more often used. As a demonstration of this general approach, consider the following differential equation and boundary conditions:

$$\frac{d^2y}{dx^2} = y^2 \tag{10.253}$$

$$y(0) = 1.0 \qquad \text{and} \qquad y(1.0) = 1.0 \tag{10.254}$$

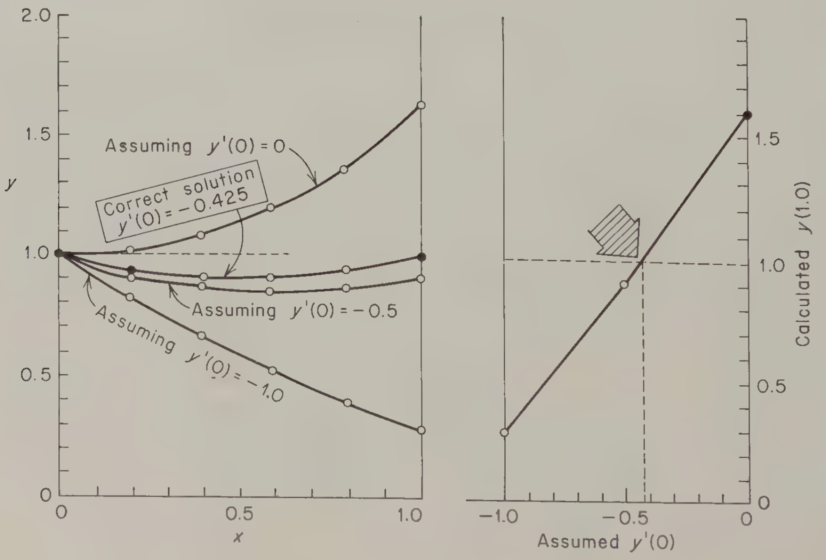
It is noted that $y'(0)$ is initially unknown. Therefore, to be able to use any of the initial-value procedures, it is necessary to first assume starting values for this derivative. The correct selection of $y'(0)$ results in $y(1) = 1.0$.

In the example, since it is being presented only for illustrative purposes, the Runge-Kutta equations listed in Table 10.10 have been used for numerical integration throughout the entire range $0 \leq x \leq 1.0$. An interval $n = 0.2$ has been presupposed as resulting in a "sufficiently accurate" solution even though such an assumption may be questioned. Function values are given in Table 10.15 for several different starting values for $y'(0)$. Corresponding curves are shown in Fig. 10.21. The correct starting value is $y'(0) = -0.425$.

Table 10.15 $y'' = y^2$, $y(0) = 1.0$, $y(1.0) = 1.0$

Function value	Assuming $y'(0) = 0$	Assuming $y'(0) = -1.0$	Assuming $y'(0) = -0.5$	Assuming $y'(0) = -0.425$
$y(0) =$	1.0000	1.0000	1.0000	1.0000
$y(0.2) =$	1.0201	0.8176	0.9188	0.9340
$y(0.4) =$	1.0821	0.6623	0.8716	0.9030
$y(0.6) =$	1.1914	0.5248	0.8550	0.9048
$y(0.8) =$	1.3581	0.3984	0.8679	0.9396
$y(1.0) =$	1.5996	0.2785	0.9111	1.0000

Figure 10.21



In general, for nonlinear boundary-value problems, *superposition does not hold.* This is the reason for the trial-and-error approach used above. However, if the problem in question is a linear one, or even if it consists of a series of linear problems, this is not necessary, and superposition of any two independent solutions allows the matching of the two specified boundary conditions. For example, if the solution corresponding to a case A is denoted as $y(x)_{\text{case } A}$, and that corresponding to case B is $y(x)_{\text{case } B}$, then a given boundary value $y(a)$ at location a can be realized by requiring that

$$C_1 y(a)_{\text{case } A} + C_2 y(a)_{\text{case } B} = y(a) \tag{10.255}$$

Similarly, at location b, the other specified boundary location,

$$C_1 y(b)_{\text{case } A} + C_2 y(b)_{\text{case } B} = y(b) \tag{10.256}$$

Casting these in matrix form yields

$$\begin{bmatrix} y(a)_{\text{case } A} & y(a)_{\text{case } B} \\ y(b)_{\text{case } A} & y(b)_{\text{case } B} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} y(a) \\ y(b) \end{bmatrix} \tag{10.257}$$

With values obtained for C_1 and C_2 from (10.257), the final solution over the entire range of interest is

$$y(x) = C_1 y(x)_{\text{case } A} + C_2 y(x)_{\text{case } B} \tag{10.258}$$

Nonlinear problems can be reduced to linear ones in a variety of ways, the accuracy of solution depending on the method of reduction used, the numerical-integration process selected, and the interval of integration chosen. For example, in the problem just solved by the trial-and-error approach, it can be assumed that the given nonlinear equation

$$y'' = y^2 \qquad (0 \leq x \leq 1.0) \tag{10.259}$$

can be represented by a series of five linear differential equations of the type

$$y'' = y_i^2 \tag{10.260}$$

where y_i is presumed to be constant over each of the intervals in question and has a magnitude equal to the beginning value of y in that interval, as shown in the following table.

	Range of solution	Differential equation
First interval	$0 \leq x \leq 0.2$	$y'' = [y(0)]^2$
Second interval	$0.2 \leq x \leq 0.4$	$y'' = [y(0.2)]^2$
Third interval	$0.4 \leq x \leq 0.6$	$y'' = [y(0.4)]^2$
Fourth interval	$0.6 \leq x \leq 0.8$	$y'' = [y(0.6)]^2$
Fifth interval	$0.8 \leq x \leq 1.0$	$y'' = [y(0.8)]^2$

With the reduction assumption and the following selected two independent initial values,

$$\begin{aligned} &\text{Case } A\text{:} \qquad y(0) = 1.0000 \qquad y'(0) = 0 \\ &\text{Case } B\text{:} \qquad y(0) = 0 \qquad y'(0) = 1.0000 \end{aligned} \tag{10.261}$$

the Runge-Kutta integration solutions shown in Table 10.16 are obtained. Here, again, it is to be recognized that the question of the applicability (or desirability) of the Runge-Kutta method as contrasted with, for example, the predictor-corrector approach over the entire range $0 \leq x \leq 1.0$ has not been considered.

Table 10.16

	Case A		Case B	
x	$y(x)$	$y'(x)$	$y(x)$	$y'(x)$
0	1.0000	0	0	1.0000
0.2	1.0200	0.2000	0.2000	1.0000
0.4	1.0808	0.4081	0.4008	1.0080
0.6	1.1858	0.6417	0.6056	1.0401
0.8	1.3422	0.9229	0.8209	1.1135
1.0	1.5628	1.2832	1.0571	1.2483

To satisfy the boundary conditions listed in Eqs. (10.261), it is required that

$$\begin{bmatrix} 1.0000 & 0 \\ 1.5628 & 1.0571 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} \tag{10.262}$$

This gives $C_1 = 1.0000$ and $C_2 = -0.5324$. The solution $y(x)$ corresponding to this reduction assumption is, therefore,

$$y(x) = 1.0000[y(x)_{\text{case } A}] - 0.5324[y(x)_{\text{case } B}] \tag{10.262a}$$

The final values are shown in Table 10.17. Also listed are the values previously obtained by the trial-and-error approach. (It is to be noted that the Runge-Kutta equations listed in Table 10.10 were used for both sets of integrations.)

There are ways other than that illustrated above to linearize nonlinear problems. One, based on a generalization of the Newton-Raphson technique for obtaining roots of nonlinear algebraic equations,† has proved to

---

† R. McGill and P. Kenneth, A Convergence Theorem on the Iterative Solution of Non-linear Two-point Boundary-value Problems, *XIV Intern. Astronaut. Federation Congr., Paris, September, 1963*, pp. 173–188.

R. McGill and P. Kenneth, Solution of Variational Problems by Means of a Generalized Newton-Raphson Operator, *AIAA J.*, vol. 2, no. 10, pp. 1761–1766, October, 1964.

Table 10.17

x	Solution of linear problem [Eq. (10.262*a*)]	Solution of nonlinear problem (Table 10.15)
0	1.0000	1.0000
0.2	0.9135	0.9340
0.4	0.8674	0.9030
0.6	0.8634	0.9048
0.8	0.9052	0.9396
1.0	1.0000	1.0000

be adaptable to a large variety of nonlinear differential-equation problems. In general, it is an iterative method. For a two-variable system, one independent and one dependent, again considering the second-order equation examined above [$y'' = f(x,y)$], the linearization is according to the following expression:

Given nonlinear equation:

$$y'' = f(x,y) \tag{10.263}$$

Corresponding linear equation:

$$({}_{n+1}y)'' = {}_n\left(\frac{\partial f}{\partial y}\right)({}_{n+1}y - {}_ny) + {}_nf(x,y) \tag{10.264}$$

It is to be noted that Eq. (10.264), in the rearranged form

$$_{n+1}y = {}_ny + \frac{({}_{n+1}y)'' - {}_nf(x,y)}{{}_n(\partial f/\partial y)} \tag{10.265}$$

corresponds to the first-order Newton equation (8.45). However, instead of convergence to a zero solution, as required in that case, convergence to $({}_{n+1}y)''$ is here presumed.

In both equations (10.264) and (10.265), the pre-subscripts indicate iterations. It is assumed, at any one stage of integration, that the n prescripted terms are known (that is, they are, for any particular integrations, constants) and that the variables in the desired differential equation are the $n + 1$ pre-subscripted terms. The differential equation is

$$({}_{n+1}y)'' = {}_{n+1}y\underbrace{\left[{}_n\left(\frac{\partial f}{\partial y}\right)\right]}_{\text{Constant}} \underbrace{- {}_ny\left[{}_n\left(\frac{\partial f}{\partial y}\right)\right] + {}_nf(x,y)}_{\text{Constant}} \tag{10.266}$$

Consider, for illustration, the formulation of the linear system corresponding to the nonlinear problem just discussed.

$$y'' = f(x,y) = y^2 \tag{10.267}$$

Since

$$\frac{\partial f}{\partial y} = 2y \tag{10.268}$$

the equation corresponding to Eq. (10.266) is

$$({}_{n+1}y)'' = {}_{n+1}y[2({}_{n}y)] - {}_{n}y[2({}_{n}y)] + ({}_{n}y)^2$$

or

$$({}_{n+1}y)'' = {}_{n+1}[2({}_{n}y)] - ({}_{n}y)^2 \tag{10.269}$$

If it is assumed as a starting situation that ${}_0y(x_i) = 1.0000$, $x_i = 0, 0.2, 0.4, 0.6, 0.8$, and 1.0, the linear differential equation (in all ranges) is

$$({}_1y)'' = {}_1y(2) - 1.0000 \tag{10.270}$$

For assumed initial values $y(0) = +1.0000$ and $y'(0) = 0$, integration using the Runge-Kutta equations of Table 10.10 results in the "new" values for ${}_1y(x)$ shown in the following table.

x	${}_0y(x)$	${}_1y(x)$
0	1.0000	1.0000
0.2	1.0000	1.0201
0.4	1.0000	1.0821
0.6	1.0000	1.1910
0.8	1.0000	1.3555
1.0	1.0000	1.5890

It is to be recognized that the linear differential equations for the next iteration are different for each range of interest.

$$({}_2y)'' = \begin{cases} ({}_2y)(1) - 1.0000 & \text{for} \quad 0 \le x \le 0.2 \\ ({}_2y)(1.0201) - 1.0406 & \text{for} \quad 0.2 \le x \le 0.4 \\ ({}_2y)(1.0821) - 1.1709 & \text{for} \quad 0.4 \le x \le 0.6 \\ ({}_2y)(1.1910) - 1.4185 & \text{for} \quad 0.6 \le x \le 0.8 \\ ({}_2y)(1.3555) - 1.8373 & \text{for} \quad 0.8 \le x \le 1.0 \end{cases} \tag{10.271}$$

Solutions for several iterations are given in Table 10.18.

For a second independent case, where it is required that the initial conditions be

$$y(0) = 0 \qquad \text{and} \qquad y'(0) = +1.0000$$

the solution is that given in Table 10.19.

Table 10.18

x	${}_0y(x)$	${}_1y(x)$	${}_2y(x)$	${}_3y(x)$
0	1.0000	1.0000	1.0000	1.0000
0.2	1.0000	1.0201	1.0201	1.0201
0.4	1.0000	1.0821	1.0821	1.0821
0.6	1.0000	1.1910	1.1914	1.1914
0.8	1.0000	1.3555	1.3579	1.3579
1.0	1.0000	1.5890	1.5987	1.5987

Table 10.19

x	${}_0y(x)$	${}_1y(x)$	${}_2y(x)$	${}_3y(x)$
0	0	0	0	0
0.2	0.2000	0.2000	0.2000	0.2000
0.4	0.4000	0.4013	0.4013	0.4013
0.6	0.6000	0.6088	0.6088	0.6088
0.8	0.8000	0.8309	0.8310	0.8310
1.0	1.0000	1.0803	1.0867	1.0867

Given the information defined in Tables 10.18 and 10.19, solution to any given boundary-value situation proceeds as defined earlier, according to Eqs. (10.257) and (10.258).

It is to be observed that in both of the cases examined, convergence is rapid. This, however, is not an isolated occurrence; rather, it is a characteristic of the procedure. In discussing the convergence of the method, McGill and Kenneth state the following: "The technique discussed herein is systematic and rapidly convergent, and owing to the quadratic convergence, great care is not required in obtaining starting functions. It is sufficient that the trial function be not so bad as to cause divergence."†

It is evident that the iterative scheme just described can be extended to the solution of other types of equations. For example, if the second derivative is defined to be also a function of y' the equation is:
Given nonlinear equation:

$$y'' = f(x,y,y') \tag{10.272}$$

---

† R. McGill and P. Kenneth, A Convergence Theorem on the Iterative Solution of Non-linear Two-point Boundary-value Problems, *XIV Intern. Astronaut. Federation Congr., Paris, September, 1963*, p. 185.

Corresponding linear equation:

$$({}_{n+1}y)'' = {}_n\left(\frac{\partial f}{\partial y}\right)({}_{n+1}y - {}_ny) + {}_n\left(\frac{\partial f}{\partial y'}\right)({}_{n+1}y' - {}_ny') + {}_nf(x,y,y') \tag{10.273}$$

If the equation is a nonlinear first-order equation, the iterative scheme is defined by Eq. (10.275).

Given nonlinear equation:

$$y' = f(x,y) \tag{10.274}$$

Corresponding linear equation:

$$({}_{n+1}y)' = {}_n\left(\frac{\partial f}{\partial y}\right)({}_{n+1}y - {}_ny) + {}_nf(x,y) \tag{10.275}$$

The generalized Newton-Raphson operator approach has a further advantage. It can be readily extended to the solution of sets of nonlinear problems. For illustration of how this might be accomplished, it is assumed that all the equations in question are of the first order, and that the independent variable is time t. The dependent variables are $x_1, x_2, x_3, \ldots, x_N$, and differentiation with respect to time is denoted by a dot over the variable in question.

In matrix form, the set of equations under consideration is given by

$$[\dot{X}] = [F(X,t)] \tag{10.276}$$

where

$$\begin{aligned} X &= x_1, x_2, x_3, \ldots, x_N \\ F &= f_1, f_2, f_3, \ldots, f_N \\ f_i &= f_i(x_1,x_2,x_3, \ldots, x_N) \end{aligned} \tag{10.277}$$

The range of independent variable of interest is

$$t_a \le t \le t_b \tag{10.278}$$

For this system, N boundary conditions are specified, part at $t = t_a$ and the remainder at t_b. For illustration, it is assumed that half are specified at t_a and the remaining $\frac{1}{2}N$ are specified at t_b; but this need not be true. The specific values are defined as follows:

$$\begin{aligned} x_1(t_a) &= x_{1a} & x_1(t_b) &= x_{1b} \\ x_2(t_a) &= x_{2a} & x_2(t_b) &= x_{2b} \\ &\cdots\cdots & &\cdots\cdots \\ x_{N/2}(t_a) &= x_{(N/2)a} & x_{N/2}(t_b) &= x_{(N/2)b} \end{aligned} \tag{10.279}$$

The linear iterative differential equation corresponding to Eqs. (10.276) is

$$[{}_{n+1}\dot{X}] = [{}_nJ(X,t)][{}_{n+1}X - {}_nX] + [{}_nF(X,t)] \tag{10.280}$$

where $[J(X,t)]$ is the jacobian matrix of partial derivatives of f_i with respect to x_j, $i = 1, 2, \ldots, N; j = 1, 2, \ldots, N$.

$$[{}_nJ(X,t)] = \begin{bmatrix} {}_n\left(\frac{\partial f_1}{\partial x_1}\right) & {}_n\left(\frac{\partial f_1}{\partial x_2}\right) & \cdots & {}_n\left(\frac{\partial f_1}{\partial x_N}\right) \\ {}_n\left(\frac{\partial f_2}{\partial x_1}\right) & {}_n\left(\frac{\partial f_2}{\partial x_2}\right) & \cdots & {}_n\left(\frac{\partial f_2}{\partial x_N}\right) \\ \cdots & \cdots & \cdots & \cdots \\ {}_n\left(\frac{\partial f_N}{\partial x_1}\right) & {}_n\left(\frac{\partial f_N}{\partial x_2}\right) & \cdots & {}_n\left(\frac{\partial f_N}{\partial x_N}\right) \end{bmatrix} \tag{10.281}$$

For a 4 × 4 set, the expanded equations are:
Given nonlinear equations:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1,x_2,x_3,x_4,t) \\ \dot{x}_2 &= f_2(x_1,x_2,x_3,x_4,t) \\ \dot{x}_3 &= f_3(x_1,x_2,x_3,x_4,t) \\ \dot{x}_4 &= f_4(x_1,x_2,x_3,x_4,t) \end{aligned} \tag{10.282}$$

Corresponding linear equations:

$$\begin{bmatrix} {}_{n+1}\dot{x}_1 \\ {}_{n+1}\dot{x}_2 \\ {}_{n+1}\dot{x}_3 \\ {}_{n+1}\dot{x}_4 \end{bmatrix} = \begin{bmatrix} {}_n\left(\frac{\partial f_1}{\partial x_1}\right) & {}_n\left(\frac{\partial f_1}{\partial x_2}\right) & {}_n\left(\frac{\partial f_1}{\partial x_3}\right) & {}_n\left(\frac{\partial f_1}{\partial x_4}\right) \\ {}_n\left(\frac{\partial f_2}{\partial x_1}\right) & {}_n\left(\frac{\partial f_2}{\partial x_2}\right) & {}_n\left(\frac{\partial f_2}{\partial x_3}\right) & {}_n\left(\frac{\partial f_2}{\partial x_4}\right) \\ {}_n\left(\frac{\partial f_3}{\partial x_1}\right) & {}_n\left(\frac{\partial f_3}{\partial x_2}\right) & {}_n\left(\frac{\partial f_3}{\partial x_3}\right) & {}_n\left(\frac{\partial f_3}{\partial x_4}\right) \\ {}_n\left(\frac{\partial f_4}{\partial x_1}\right) & {}_n\left(\frac{\partial f_4}{\partial x_2}\right) & {}_n\left(\frac{\partial f_4}{\partial x_3}\right) & {}_n\left(\frac{\partial f_4}{\partial x_4}\right) \end{bmatrix} \begin{bmatrix} {}_{n+1}x_1 - {}_nx_1 \\ {}_{n+1}x_2 - {}_nx_2 \\ {}_{n+1}x_3 - {}_nx_3 \\ {}_{n+1}x_4 - {}_nx_4 \end{bmatrix} + \begin{bmatrix} {}_nf_1(x_1,x_2,x_3,x_4,t) \\ {}_nf_2(x_1,x_2,x_3,x_4,t) \\ {}_nf_3(x_1,x_2,x_3,x_4,t) \\ {}_nf_4(x_1,x_2,x_3,x_4,t) \end{bmatrix} \tag{10.283}$$

The first of these linear expressions is

$$\begin{aligned} {}_{n+1}\dot{x}_1 = {}& {}_n\left(\frac{\partial f_1}{\partial x_1}\right)({}_{n+1}x_1 - {}_nx_1) + {}_n\left(\frac{\partial f_1}{\partial x_2}\right)({}_{n+1}x_2 - {}_nx_2) \\ &+ {}_n\left(\frac{\partial f_1}{\partial x_3}\right)({}_{n+1}x_3 - {}_nx_3) + {}_n\left(\frac{\partial f_1}{\partial x_4}\right)({}_{n+1}x_4 - {}_nx_4) + {}_nf_1(x_1,x_2,x_3,x_4,t) \end{aligned} \tag{10.284}$$

It is to be remembered that the n pre-subscripted terms represent, for any one cycle of iteration, constants.

The boundary conditions for the system of equations under consideration are, for example,

$$x_1(t_a) = x_{1a} \qquad x_1(t_b) = x_1 b$$

and

$$x_2(t_a) = x_{2a} \qquad x_2(t_b) = x_2 b \tag{10.285}$$

This number of constraints is consistent with the prescribed four-dependent-variable first-order systems. Correspondingly, to be able to realize these specified values x_{1a}, x_{2a}, x_{1b}, and x_{2b} by using the method of superposition, four independent solutions must be obtained. In selecting these independent cases it is to be recognized that certain assumptions for the initial conditions can facilitate the numerical process of solving the resulting linear algebraic equations. Precision can also be improved by judicious selection of these values.

It is to be recognized that the linear differential equations (10.283) are equivalent to

$$[\dot{X}(t)] = [A(t)][X(t)] + [B(t)] \tag{10.286}$$

However, not all cases need be solved by using this full equation. If, for example, only the influence of $x_3(t_a)$ were desired, it would be sufficient to examine the solution of the homogeneous equation

$$[\dot{X}(t)] = [A(t)][X(t)] \tag{10.287}$$

For the four-variable system, then, initial conditions, as shown in Figs. 10.22 to 10.25, might be specified.

Figure 10.22

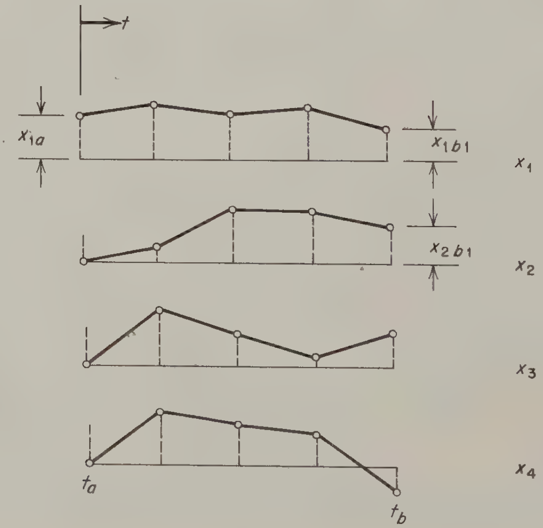
Case I

$$x_1(t_a)_{\text{I}} = x_{1a}$$

$$x_2(t_a)_{\text{I}} = 0$$

$$x_3(t_a)_{\text{I}} = 0$$

$$x_4(t_a)_{\text{I}} = 0$$



Case II

$$x_1(t_a)_{\text{II}} = 0$$

$$x_2(t_a)_{\text{II}} = x_{2a}$$

$$x_3(t_a)_{\text{II}} = 0$$

$$x_4(t_a)_{\text{II}} = 0$$

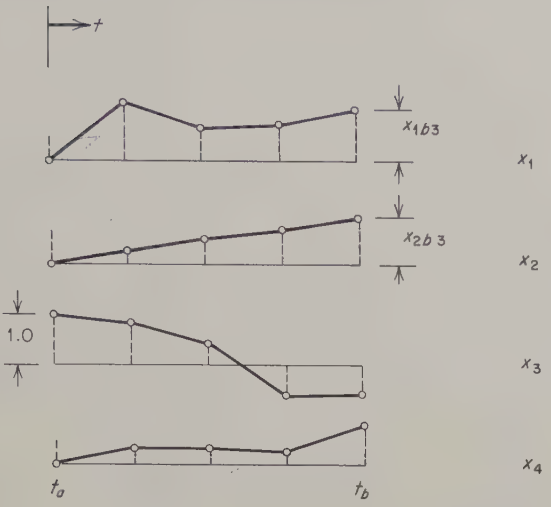


Figure 10.23

Under these assumptions, the algebraic equations to be solved to allow matching of the boundary conditions are

$$\begin{bmatrix} x_{1a} & 0 & 0 & 0 \\ 0 & x_{2a} & 0 & 0 \\ x_{1b1} & x_{1b2} & x_{1b3} & x_{1b4} \\ x_{2b1} & x_{2b2} & x_{2b3} & x_{2b4} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{bmatrix} = \begin{bmatrix} x_{1a} \\ x_{2a} \\ x_{1b} \\ x_{2b} \end{bmatrix} \tag{10.288}$$

Case III

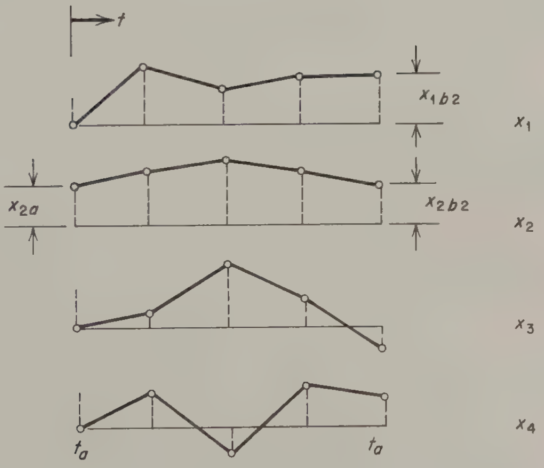
$$x_1(t_a)_{\text{III}} = 0$$

$$x_2(t_a)_{\text{III}} = 0$$

$$x_3(t_a)_{\text{III}} = 1.0$$

$$x_4(t_a)_{\text{III}} = 0$$

Figure 10.24



CASE IV

$$x_1(t_a)_{\mathrm{IV}} = 0$$

$$x_2(t_a)_{\mathrm{IV}} = 0$$

$$x_3(t_a)_{\mathrm{IV}} = 0$$

$$x_4(t_a)_{\mathrm{IV}} = 1.0$$

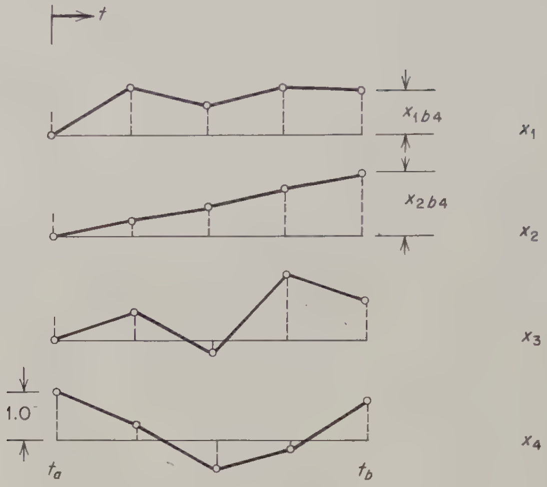


Figure 10.25

The C's having been determined, the final expression for $x_i(t)$ is

$$x_i(t) = C_1(x_i)_{\mathrm{I}} + C_2(x_i)_{\mathrm{II}} + C_3(x_i)_{\mathrm{III}} + C_4(x_i)_{\mathrm{IV}} \qquad (i = 1, 2, 3, 4) \tag{10.289}$$

For a numerical example, consider again the solution to the system

$$y'' = y^2 \tag{10.290}$$

and $\quad y(0) = 1.0 \quad$ and $\quad y(1.0) = 1.0 \quad$ (10.291)

By substituting $z = y'$, the second-order nonlinear equation can be reduced to two first-order equations:

$$y' = z \tag{10.292}$$

$$z' = y^2 \tag{10.293}$$

The boundary conditions remain the same as those listed in Eq. (10.291).

The first-order linear iteration formula corresponding to Eq. (10.292) is

$${}_{n+1}y' = 1({}_{n+1}z - {}_nz) + {}_nz = {}_{n+1}z \tag{10.294}$$

The second equation, corresponding to Eq. (10.293), is

$${}_{n+1}z' = 2\,{}_ny({}_{n+1}y - {}_ny) + {}_ny^2 = 2({}_ny)({}_{n+1}y) - ({}_ny)^2 \tag{10.295}$$

By using the Runge-Kutta equations of Table 10.8, assuming as initial conditions $y(0) = 1.0$ and $z(0) = 0$, the solutions listed in Table 10.20 were obtained. It should be noted that because of the particular nature of the linear equation (10.292), it was not necessary to make an initial assumption for $z(x)$.

Table 10.20 $y' = z,\ z' = y^2,\ y(0) = 1.0,\ z(0) = 0$

Function values	Initial assumption	First iteration	Second iteration	Third iteration
$y(0) =$	1.0000	1.0000	1.0000	1.0000
$y(0.2) =$	1.0000	1.0201	1.0201	1.0201
$y(0.4) =$	1.0000	1.0821	1.0821	1.0821
$y(0.6) =$	1.0000	1.1910	1.1913	1.1913
$y(0.8) =$	1.0000	1.3556	1.3577	1.3577
$y(1.0) =$	1.0000	1.5891	1.5893	1.5893
$z(0) =$		0	0	0
$z(0.2) =$		0.2027	0.2027	0.2027
$z(0.4) =$		0.4217	0.4220	0.4220
$z(0.6) =$		0.6746	0.6780	0.6780
$z(0.8) =$		0.9819	0.9990	0.9989
$z(1.0) =$		1.3682	1.4290	1.4290

Prob. 10.1: Using at least two of the methods discussed in this chapter, solve the following initial-value problems, assuming $h = 0.2$:

(*a*) $y' = 1 - x^2y;\ y(0) = 1.0;\ 0 \leq x \leq 1.0$
(*b*) $y'' = 1 - xy;\ y(0) = 1.0,\ y'(0) = 1.0;\ 0 \leq x \leq 1.0$
(*c*) $(y'')(y') + 1 = 0;\ y(0) = 0,\ y'(0) = 1.0;\ 0 \leq x \leq 2.0$
(*d*) $y'' + A \sin y = 0;\ y(0) = a,\ y'(0) = 0;\ 0 \leq x \leq 1.0;$
where $A = 2.0,\ a = 0.5$
(*e*) $\theta' + 2\theta + 0.1\theta^2 = 10;\ \theta(0) = 0;\ 0 \leq \theta \leq 1.0$
(*f*) $z'' + 40(1 + \cos x)z = \cos 3x;\ z(0) = 0,\ z'(0) = 0;\ 0 \leq x \leq 1.0$
(*g*) $\ddot{x} + 0.1x + 1/x^2 = 0;\ x(0) = 1.0,\ \dot{x}(0) = 10.0;\ 0 \leq t \leq 2.0$
(*h*) $F'' = F' - FF' + 4;\ F(0) = 0,\ F'(0) = 0;\ 0 \leq x \leq 1.0$
(*i*) $B''' - \lambda^2B' + B + C = 0;\ B''(0) = 0,\ B'(0) = 1.0,\ B(0) = 1.0;$
$0 \leq x \leq 1.0;\ C = 3.0$

Prob. 10.2: To three significant figures, determine the solution to the following sets of differential equations:

(*a*) $\dot{x} = x + y + t$
$\dot{y} = 2x + y$ $\qquad x(0) = 0,\ y(0) = 1.0;\ 0 \leq t \leq 0.5$

(*b*) $\dot{x} = x + y + t$
$\ddot{y} = x - t^2$ $\qquad x(0) = 0,\ y(0) = 1.0,\ \dot{y}(0) = 0;\ 0 \leq t \leq 1.0$

(*c*) $\dot{x} = x + 2y + t$
$\dot{y} = y + 2z + t$ $\qquad x(0) = 0,\ y(0) = 1.0,\ z(0) = 1.0;\ 0 \leq t \leq 1.0$
$\dot{z} = z + 2x + t$

(*d*) $\ddot{x} = y + e^x$
$\dot{y} = x + t$ $\qquad x(0) = 1.0,\ \dot{x}(0) = 0,\ y(0) = 0;\ 0 \leq t \leq 1.0$

Prob. 10.3: Using the predictor-corrector method of Milne, extend the range of solution of Table 10.13 for the equation $y' = 1 - xy$, from that given ($x = 0$, 0.2, 0.4, and 0.6) to $x = 2.0$. Use an interval of $h = 0.2$.

Prob. 10.4: For the given set of equations

$$\begin{aligned} \dot{x} &= 3x + y + t \\ \dot{y} &= x - 2t \end{aligned} \qquad x(0) = 1.0,\ y(0) = -1.0$$

obtain starting values at $t = 0$, 0.2, 0.4, and 0.6. Then, using the Milne predictor-corrector method, extend the range of solution to $t = 2.0$. The increment should be selected as $h = 0.2$, and four significant figures are desired.

Prob. 10.5: Determine values of y at four intermediate pivotal locations in the range $0 \leq x \leq 1.0$, for each of the following boundary-value problems:

(*a*) $y'' + 4y = 2x + 1;\ y(0) = 0,\ y(1.0) = 1.0$
(*b*) $y'' + 4y = e^x;\ y(0) = 0,\ y(0.75) = 1.0$
(*c*) $y^{\mathrm{IV}} + k^2y'' + 4 = 0;\ y(0) = 0,\ y''(0) = 0,\ y(1.0) = 0,\ y'(1.0) = 0$
(*d*) $y'' - Cy' + 4 = 0;\ y(0) = 1.0,\ y(1.0) = 0$

Prob. 10.6: For the homogeneous equation $y'' + \lambda^2 y = 0$, determine the first two eigenvalues, when the boundary conditions are

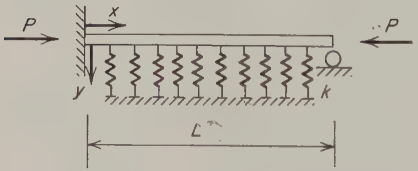
(*a*) $y(0) = 0,\ y(1.0) = 0$
(*b*) $y'(0) = 0,\ y(1.0) = 0$
(*c*) $y(0) = 0,\ y'(1.0) = 0$
(*d*) $y'(0) = 0,\ y'(1.0) = 0$

Prob. 10.7: An elastically supported, uniform cross-section column of length L, fixed at one end and pin-connected at the other, is subjected to an axial thrust P, as shown in Fig. 10.26. The governing differential equation is

$$\frac{d^4y}{dx^4} + \frac{P}{EI}\frac{d^2y}{dx^2} - \frac{k}{EI}y = 0$$

where k is the spring constant of the elastic support (force per unit of deflection per unit of length along the member), and EI represents the

Figure 10.26



bending stiffness per unit of length of the cross section. The boundary conditions are

$$y(0) = 0 \qquad y'(0) = 0$$
$$y(L) = 0 \qquad y''(L) = 0$$

To change the range of integration, it is assumed that a new independent variable z is introduced, where

$$z = \frac{x}{L}$$

The resulting equation is

$$\frac{d^4y}{dz^4} + \frac{PL^2}{EI}\frac{d^2y}{dz^2} - \frac{kL^4}{EI}y = 0$$

The corresponding boundary conditions are

$$y(0) = 0 \qquad y'(0) = 0$$
$$y(1.0) = 0 \qquad y''(1.0) = 0$$

(*a*) For the case where $\gamma = kL^4/PL^2 = 2.5$, determine the first two eigenvalues and their corresponding eigenvectors.

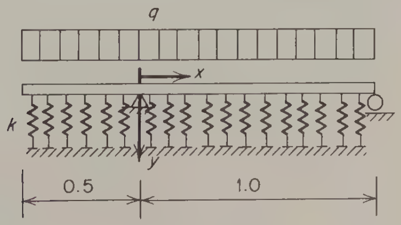
(*b*) For the case where $kL^4/EI = 5.0$, determine the first two eigenvalues and their corresponding eigenvectors.

Prob. 10.8: The elastically restrained, uniform cross-section beam shown (Fig. 10.27) is subjected to a uniform lateral load of q lb per unit of length. The governing differential equation is

$$\frac{d^4y}{dx^4} + \frac{kL^4}{EI}y = \frac{qL^4}{EI}$$

where k is the spring constant of the foundation. The boundary condi-

Figure 10.27



tions are specified to be

$$y''(-0.5) = 0 \qquad y'''(-0.5) = 0$$

$$y(0) = 0$$

$$y(+1.0) = 0 \qquad y''(+1.0) = 0$$

Using the collocation method ($h = 0.25$), determine the deflection pattern along the length of the beam, assuming $kL^4/EI = 230$ and $qL^4/EI = 300$.

Prob. 10.9: Solve the following nonlinear initial-value problems for the range of independent variable indicated:

(*a*) $(f'')^2 - \frac{2}{3}(f' - 1)^2(f' + 2) = 2$
$f(0) = 1.0;\ 0 \le x \le 1.0$
$f'(0) = 0;\ h = 0.05$

(*b*) $\ddot{y} = u\dot{y} - uy^2(\dot{y}) - y;\ u = 0.95$
$y(0) = 1.4;\ 0 \le t \le 1.0$
$y'(0) = 0;\ h = 0.05$

(*Note:* For these problems, use the Kutta-Simpson one-third rule for the range $0 \le t \le 0.2$, $h = 0.05$. Continue the solution by using the Milne method for $0.2 \le t \le 1.0$.)

Prob. 10.10: For the nonlinear equation

$$\phi'' + \phi\phi' + x = 0$$

the required boundary conditions are $\phi(0) = 0$ and $\phi'(\infty) = 1.0$; however, the range of primary interest is only $0 \le x \le 2.0$.
[*Note:* In solving this problem, first assume that ∞ can be considered to be 5.0. Next, obtain a second solution, assuming ∞ corresponds to 10.0. If the two solutions thus obtained are (for the range of x of interest) within the accuracy desired, the values computed in that range can be considered to be the solution. If they are not, it is necessary to carry out a third solution assuming, for example, that ∞ equals 20.0. With three or more independent solutions, obtained in the manner just described, extrapolation can be used to define the "true values."]

Prob. 10.11: A 14WF43 pin-connected steel column of 280 in. length is subjected to a doubly eccentric axial thrust of 10,000 lb. The eccentricities are $e_x = 5$ in. and $e_y = 5$ in. If u is the displacement of the shear center in the x direction, v is the displacement of the shear center in the y direction, and β is the angle of twist about the z axis, the equations governing the deformation are

$$v'' = A(-5 - v + 5\beta)$$

$$u'' = B(-5 - u - 5\beta)$$

$$\beta''' = C\beta' + D(-u' + v')$$

where
$$A = 0.778 \times 10^{-6}$$
$$B = 0.739 \times 10^{-5}$$
$$C = 1.729 \times 10^{-4}$$
$$D = 0.856 \times 10^{-6}$$

The boundary conditions are

$$u(0) = v(0) = \beta(0) = \beta''(0) = 0$$
$$u(L) = v(L) = \beta(L) = \beta''(L) = 0$$

(It is to be noted that in all equations differentiation is with respect to the independent variable z.) Using the collocation method, assuming the total length is subdivided into four equal segments, determine u, v, and β at the midpoint of the member (i.e., at $z = 140$ in.).

11 numerical solution of partial differential equations

11.1 Introduction

For many problems the dependent variable is defined in terms of more than one independent variable. Variation with respect to the independent parameters is possible, and as pointed out in earlier chapters, the derivatives describing these independent variations are defined as partial derivatives. Differential equations involving partial derivatives are known as partial differential equations.

Equation (10.3), which is here reproduced as (11.1), is a general second-order linear partial differential equation in the independent variables x and y and in the dependent variable z.

$$A_1(x,y)\frac{\partial^2 z}{\partial x^2} + A_2(x,y)\frac{\partial^2 z}{\partial x\,\partial y} + A_3(x,y)\frac{\partial^2 z}{\partial y^2} + A_4(x,y)\frac{\partial z}{\partial x} + A_5(x,y)\frac{\partial z}{\partial y} + A_6(x,y) + A_7(x,y)z = 0 \quad (11.1)$$

If this equation is rewritten in the form

$$A_1(x,y)\frac{\partial^2 z}{\partial x^2} + A_2(x,y)\frac{\partial^2 z}{\partial x\,\partial y} + A_3(x,y)\frac{\partial^2 z}{\partial y^2} + f\left(x,y,z,\frac{\partial z}{\partial x},\frac{\partial z}{\partial y}\right) = 0 \quad (11.2)$$

and if $f\left(x,y,z,\frac{\partial z}{\partial x},\frac{\partial z}{\partial y}\right)$ is allowed to be either linear or nonlinear, the partial differential equation (11.2) is classified as "quasi-linear."

In general, for the purpose of classification and study, second-order quasi-linear partial differential equations are grouped according to the nature of the conic associated with their characteristic quadratic form

$$Q(\eta,\epsilon) = A_1\eta^2 + A_2\eta\epsilon + A_3\epsilon_2 \tag{11.3}$$

The three possible cases and the necessary condition associated with each are defined in Eqs. (11.4).

$$\begin{aligned} &\text{Elliptic equation:} && A_2{}^2 - 4A_1A_3 < 0 \\ &\text{Parabolic equation:} && A_2{}^2 - 4A_1A_3 = 0 \\ &\text{Hyperbolic equation:} && A_2{}^2 - 4A_1A_3 > 0 \end{aligned} \tag{11.4}$$

Unlike ordinary differential equations, where for "real cases" a given equation may have either initial or boundary conditions associated with it, particular forms of second-order two-primary-variable partial differential equations are almost without exception connected with particular types of associated conditions. Elliptic equations, for example, are of the *equilibrium*, or *steady-state*, type with boundary conditions of the Dirichlet form. That is, the boundary conditions require that either the function, or its normal derivative, or a linear combination of these must take on certain specified values on the *closed boundary*, within which boundary the solution is desired. On the other hand, parabolic- and hyperbolic-type equations are for *propagation* or *non-steady-state* situations. For both of these types of equations, solutions are not restricted to a *closed region* but rather are allowed to propagate in an *open domain*. It must be understood, however, that this is not to say that for these types of problems conditions associated with certain of the variables cannot be of the boundary type. The distinction is that at least one is in an open domain.

The torsional behavior of shafts, the potential distribution in a transformer coil, and the potential flow of fluids are all examples of steady-state boundary-value-type engineering problems that require solution of the following elliptic partial differential equations:

$$\text{Laplace's equation†} \qquad \nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0 \tag{11.5}$$

$$\text{Poisson's equation:} \qquad \nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = f(x,y) \tag{11.6}$$

† ∇^2 is pronounced "del squared" and is frequently referred to as the *Laplace operator*.

The one-dimensional heat-flow equation (11.7) is an example of a parabolic-type partial differential equation which is often encountered.

$$K \frac{\partial^2 \tau}{\partial x^2} = \frac{\partial \tau}{\partial t} \tag{11.7}$$

The one-dimensional wave equation (11.8), which can be used to define the vibration of a string, is a hyperbolic-type partial differential equation.

$$C^2 \frac{\partial^2 \phi}{\partial x^2} = \frac{\partial^2 \phi}{\partial t^2} \tag{11.8}$$

It should be recognized that the problem types defined by Eqs. (11.4) depend on the magnitudes of A_1, A_2, and A_3. These quantities, however, may be specified as functions of x and y. It is therefore possible for a given equation [of the (11.2) type] to be elliptic in one range of independent variable, parabolic in a second range, and hyperbolic in still a third.

While the discussion thus far has been restricted to second-order equations of two independent variables, partial differential equations of other orders do exist, and have meaning. For example, the lateral deformation of an elastic plate is a steady-state (or equilibrium) type of problem, whose solution is governed by the biharmonic equation†

$$\nabla^4 \phi = \nabla^2(\nabla^2 \phi) = \frac{\partial^4 \phi}{\partial x^4} + 2 \frac{\partial^4 \phi}{\partial x^2\, dy^2} + \frac{\partial^4 \phi}{\partial y^4} = F(x,y) \tag{11.9}$$

The two-dimensional heat-flow equation is

$$\frac{\partial^2 \tau}{\partial x^2} + \frac{\partial^2 \tau}{\partial y^2} = C \frac{\partial \tau}{\partial t} \tag{11.10}$$

The *ordinary finite-difference method*, referred to in Chap. 10 as the *collocation method*, provides a simple and effective, yet general, method by which reasonable, quantitative solutions may be obtained at preselected pivotal points within the domain in question. The accuracy that is achieved with this method depends on the equation given and the grid type and size selected.

11.2 Two-dimensional Grids

Depending primarily on the shape of the boundary of the region for which solution is desired, the following types of two-dimensional grids are often used:

1. Square
2. Rectangular

---

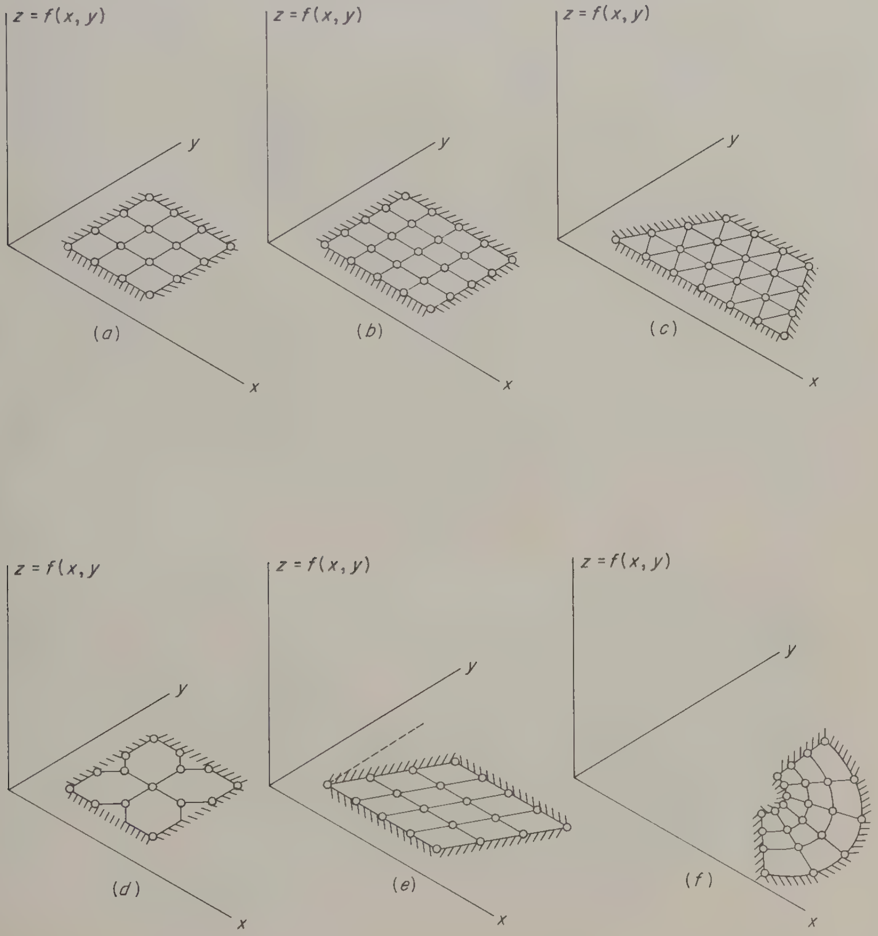
† ∇^4 is pronounced "del fourth" and is frequently referred to as the *biharmonic operator*.

3. Triangular
4. Hexagonal
5. Parallelogram (“skewed” coordinates)
6. Circular

These are illustrated in Fig. 11.1*a* through *f*, where it has been assumed that the independent variables are x and y and the dependent variable is $z(x,y)$.

In some cases, irregularly curved boundaries such as those illustrated in Fig. 11.2 are encountered. Boundaries such as these cannot be exactly represented by any of the grids shown in Fig. 11.1, and therefore special

Figure 11.1 *(a)* Square; *(b)* rectangular; *(c)* triangular; *(d)* hexagonal; *(e)* parallelogram (skewed coordinates); *(f)* circular.



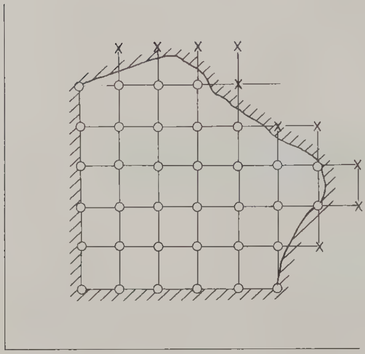
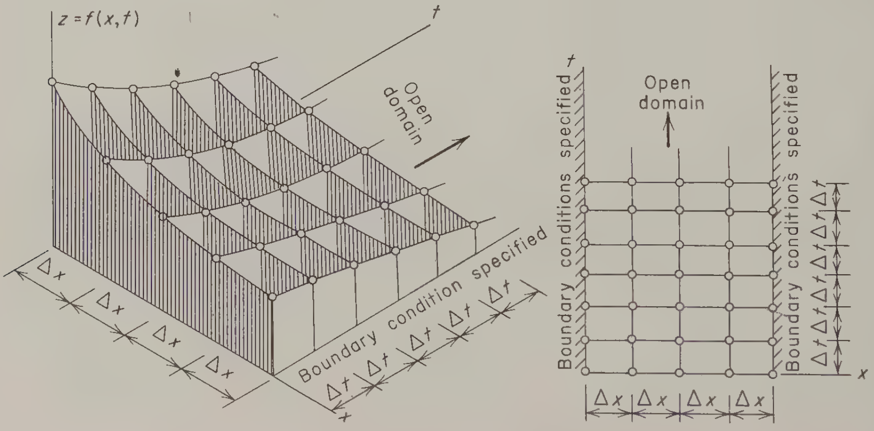


Figure 11.2

techniques are required. An example of one of the methods is discussed in Sec. 11.4.

All the grid systems shown in Fig. 11.1 are directly associated with, and therefore directly applicable to, boundary-value situations. However, as pointed out in Sec. 11.1, not all partial differential equation problems are necessarily of this type. For the case where the independent variables are x and t, t being time, and the dependent variable is $z(x,t)$, the situation probably would be more like that shown in Fig. 11.3. Here, the grid spacing in the x direction, the direction associated with the prescribed boundary conditions, is selected so that the total dimension is subdivided into an integral number of segments Δx. The greater the number of segments, the better will be the accuracy of the numerical approximation. In the t direction, however, if the problem in that direc-

Figure 11.3



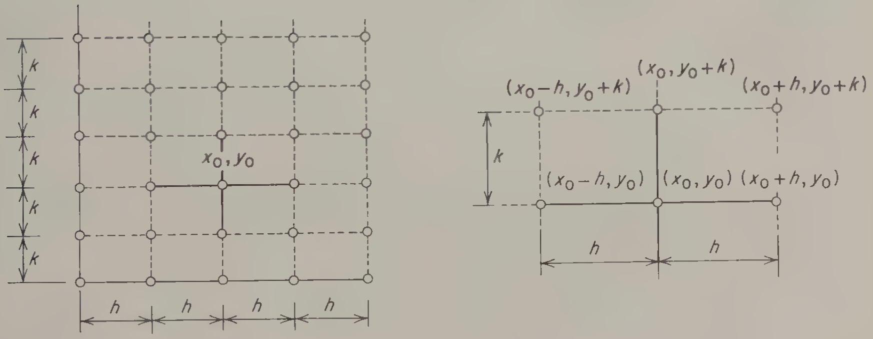
tion is of the propagation type, there are no similar "geometrical constraints" (such as division by an integral number). Δt is usually selected sufficiently small to ensure stability of the numerical process. This effect will be demonstrated later in this chapter.

11.3 Development of Finite-difference Computational Molecules

It will be recalled that the primary advantage realized in using the collocation method in the solution of ordinary differential equations was the ease with which the influence of the various terms in a given equation could be taken into account. The difference operators [for example, $D^2(y)$] were initially treated as independent quantities, and their individual influences were cataloged. (See, for example, Table 9.27.) The composite effect of the various terms in a given differential equation was then taken into account by simple algebraic means using these pre-established relationships. This was possible because in each instance operations were written in terms of the function values at the preselected pivotal locations. This same procedure can be extended to the solution of partial differential equations. It should be recognized, however, that by the very nature of a multi-independent-variable situation as opposed to the one-independent-variable case considered before, cataloging will not be as convenient.

To illustrate the general concept, it will be assumed that the system in question has two independent variables x and y. The dependent variable will be presupposed to be $\phi(x,y)$. A rectangular grid of the type illustrated in Fig. 11.4, where the spacing in the x direction is h and that in the y direction is k, has been selected. It should be noted that the coordinates

Figure 11.4



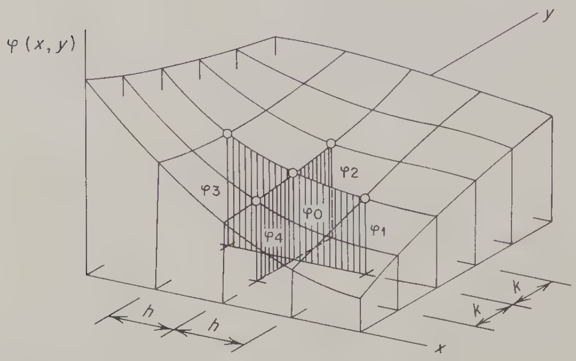


Figure 11.5

of any pivotal location can be defined in terms of a given reference location (x_0,y_0) by the expressions

$$x = x_0 \pm ih \qquad y = y_0 \pm jk \tag{11.11}$$

where i and j are integers.

It will be recalled that in Chap. 10, in the solution of ordinary differential equations by the collocation method, it was convenient for reference purposes to number the various pivotal locations. The same is true for partial differential equations. In Fig. 11.5, five locations, centered symmetrically about location 0, are presumed to be of interest. By using the information given in Table 9.27 and assuming the acceptability of an error of the order of h^2 and expansion about the center location 0, the various "one-directional" partial derivatives are

$$D_x(\varphi) = \frac{\partial \phi}{\partial x} \cong \frac{1}{2h}(-\varphi_3 + \varphi_1) \tag{11.12}$$

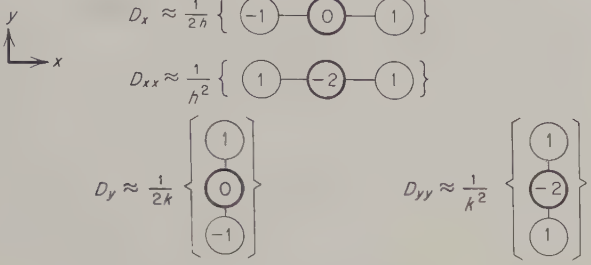
$$D_{xx}(\varphi) = \frac{\partial^2 \varphi}{\partial x^2} \cong \frac{1}{h^2}(\varphi_3 - 2\varphi_0 + \varphi_1) \tag{11.13}$$

$$D_y(\varphi) = \frac{\partial \varphi}{\partial y} \cong \frac{1}{2k}(-\varphi_4 + \varphi_2) \tag{11.14}$$

$$D_{yy}(\varphi) = \frac{\partial^2 \varphi}{\partial y^2} \cong \frac{1}{k^2}(\varphi_4 - 2\varphi_0 + \varphi_2) \tag{11.15}$$

. . . .

These equations can be represented by "stencils" or "computational molecules" of the following type:


y
x
$D_x \approx \frac{1}{2h}$
-1
0
1
$D_{xx} \approx \frac{1}{h^2}$
1
-2
1
$D_y \approx \frac{1}{2k}$
1
0
-1
$D_{yy} \approx \frac{1}{k^2}$
1
-2
1

It should be noted that in each case the location about which expansion is to occur is shown in boldface. For example, in the first illustrated case, expansion is presumed about the center location. D_x is then approximately equal to $1/2h$ times the quantity (-1 times the value of the function at the first pivotal location to the left of the reference point, plus 1 times the value of the function at the first pivotal location to the right). Note that such a computational-molecule representation of the operation $\partial/\partial x$ is independent of the particular numbering of the reference locations.

The finite-difference representation of the Laplace operator $\nabla^2\varphi$ is developed in the following fashion:

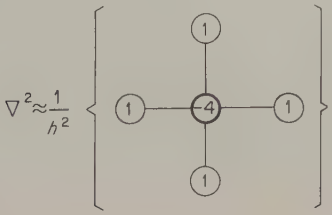
$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} \cong \frac{1}{h^2}(\varphi_3 - 2\varphi_0 + \varphi_1) + \frac{1}{k^2}(\varphi_4 - 2\varphi_0 + \varphi_2) \qquad (11.16)$$

If the grid is presumed square, that is, $k = h$, the finite-difference representation becomes

$$\nabla^2\varphi \cong \frac{1}{h^2}(\varphi_3 - 2\varphi_0 + \varphi_1 + \varphi_4 - 2\varphi_0 + \varphi_2)$$

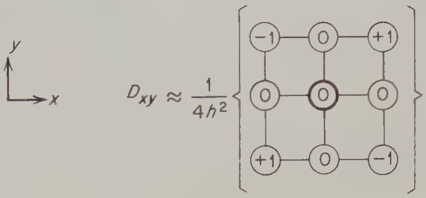
$$\cong \frac{1}{h^2}(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 - 4\varphi_0) \qquad (11.17)$$

The corresponding computational molecule is



(11.18)

The required relationship for the second mixed partial derivative $\partial^2\varphi/\partial x\, \partial y$, again assuming a square grid, is obtained by applying the operator D_x to D_y, or vice versa. In both cases the computational molecule shown in Eq. (11.19) is obtained.



$$D_{xy} \approx \frac{1}{4h^2}\left\{\begin{matrix} -1 & 0 & +1 \\ 0 & 0 & 0 \\ +1 & 0 & -1 \end{matrix}\right\} \tag{11.19}$$

A summary of the more often used square-grid computational molecules is given in Table 11.1. For those cases shown, it has been assumed that an error of the order of h^2 is permissible. The molecules were obtained by using the information contained in Table 9.27.

Square-grid higher-order computational molecules for D_{xy}, ∇^2, and ∇^4, assuming a permissible error of the order of h^4, are shown in Table 11.2. Again, the information contained in Table 9.27 was used to develop the molecules.

Two-dimensional integration formulas (i.e., volume determinations) can also be described as computational molecules. For the double integral

$$V = \int\int f(x,y)\, dx\, dy \tag{11.20}$$

Table 11.3 summarizes the trapezoidal and Simpson's one-third rule cases. As was true in all the situations illustrated in Tables 11.1 and 11.2, it has been presumed in Table 11.3 that the region in question can be subdivided into a square-grid system of dimensions $h \times h$.

11.3.1 "Skewed"-coordinate Computational Molecules

It should be evident that computational molecules can be developed also for noncartesian coordinate systems. When the spatial domains in question are parallelogrammatic in form, an often-used noncartesian system is the "skewed" one illustrated in Fig. 11.6. Here the coordinate directions u and v are at an angle α, one to the other. It should be noted that for the case illustrated, the coordinate dimensions u_i and v_i, which define the general point P_i, are measured in the u and v directions. In an xy cartesian system the corresponding situation is that shown in Fig. 11.7.

If it is assumed for a given situation that the origins 0 of the two systems

Table 11.1 Computational Molecules (Square Grid): Symmetrical Partial Finite-difference Operators (Errors of the Order of h^2)

$$D_x \approx \frac{1}{2h}\left\{\begin{matrix} -1 & 0 & +1 \end{matrix}\right\} \qquad D_y \approx \frac{1}{2h}\left\{\begin{matrix} +1 \\ 0 \\ -1 \end{matrix}\right\}; \quad D_{yy} \approx \frac{1}{h^2}\left\{\begin{matrix} +1 \\ -2 \\ +1 \end{matrix}\right\}$$

$$D_{xx} \approx \frac{1}{h^2}\left\{\begin{matrix} +1 & -2 & +1 \end{matrix}\right\}$$

$$D_{xxx} \approx \frac{1}{2h^3}\left\{\begin{matrix} -1 & +2 & 0 & -2 & +1 \end{matrix}\right\}$$

$$D_{xxxx} \approx \frac{1}{h^4}\left\{\begin{matrix} +1 & -4 & +6 & -4 & +1 \end{matrix}\right\}; \quad D_{yyy} \approx \frac{1}{2h^3}\left\{\begin{matrix} +1 \\ -2 \\ 0 \\ +2 \\ -1 \end{matrix}\right\}; \quad D_{yyyy} \approx \frac{1}{h^4}\left\{\begin{matrix} +1 \\ -4 \\ +6 \\ -4 \\ +1 \end{matrix}\right\}$$

$$D_{xy} \approx \frac{1}{4h^2}\left\{\begin{matrix} -1 & 0 & +1 \\ 0 & 0 & 0 \\ +1 & 0 & -1 \end{matrix}\right\} \qquad D_{xxy} \approx \frac{1}{2h^3}\left\{\begin{matrix} +1 & -2 & +1 \\ 0 & 0 & 0 \\ -1 & +2 & -1 \end{matrix}\right\}$$

$$D_{xxyy} \approx \frac{1}{h^4}\left\{\begin{matrix} +1 & -2 & +1 \\ -2 & +4 & -2 \\ +1 & -2 & +1 \end{matrix}\right\} \qquad D_{xyy} \approx \frac{1}{2h^3}\left\{\begin{matrix} -1 & 0 & +1 \\ +2 & 0 & -2 \\ -1 & 0 & +1 \end{matrix}\right\}$$

$$\nabla^2 \approx \frac{1}{h^2}\left\{\begin{matrix} & +1 & \\ +1 & -4 & +1 \\ & +1 & \end{matrix}\right\} \qquad \nabla^4 \approx \frac{1}{h^4}\left\{\begin{matrix} & & +1 & & \\ & +2 & -8 & +2 & \\ +1 & -8 & +20 & -8 & +1 \\ & +2 & -8 & +2 & \\ & & +1 & & \end{matrix}\right\}$$

---

shown in Figs. 11.6 and 11.7 coincide, that the directions u and x are one and the same, and that $P_i(u_i,v_i)$ and $P_i(x_i,y_i)$ define the same point, the two cases can be related as illustrated in Fig. 11.8. The transformation equations are

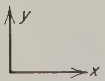
$$x = u + v \cos \alpha \tag{11.21}$$

$$y = v \sin \alpha \tag{11.22}$$

Table 11.2 Higher-order Computational Molecules (Square Grid): Symmetrical Partial Finite-difference Operators (Errors of the Order of h^4)

$$D_{xy} \approx \frac{1}{144h^2}\left\{\begin{array}{ccccc} -1 & +8 & 0 & -8 & +1 \\ +8 & -64 & 0 & +64 & -8 \\ 0 & 0 & \boxed{0} & 0 & 0 \\ -8 & +64 & 0 & -64 & +8 \\ +1 & -8 & 0 & +8 & -1 \end{array}\right\}$$

$$\nabla^2 \approx \frac{1}{12h^2}\left\{\begin{array}{ccccc} & & -1 & & \\ & & +16 & & \\ -1 & +16 & \boxed{-60} & +16 & -1 \\ & & +16 & & \\ & & -1 & & \end{array}\right\}$$



$$\nabla^4 \approx \frac{1}{6h^4}\left\{\begin{array}{ccccccc} & & & -1 & & & \\ & & -1 & +14 & -1 & & \\ & -1 & +20 & -77 & +20 & -1 & \\ -1 & +14 & -77 & \boxed{+184} & -77 & +14 & -1 \\ & -1 & +20 & -77 & +20 & -1 & \\ & & -1 & +14 & -1 & & \\ & & & -1 & & & \end{array}\right\}$$

(Because of the general nature of the process, the subscripts have not been included.) Independent variations of x and y with respect to u and v result in the partial derivatives

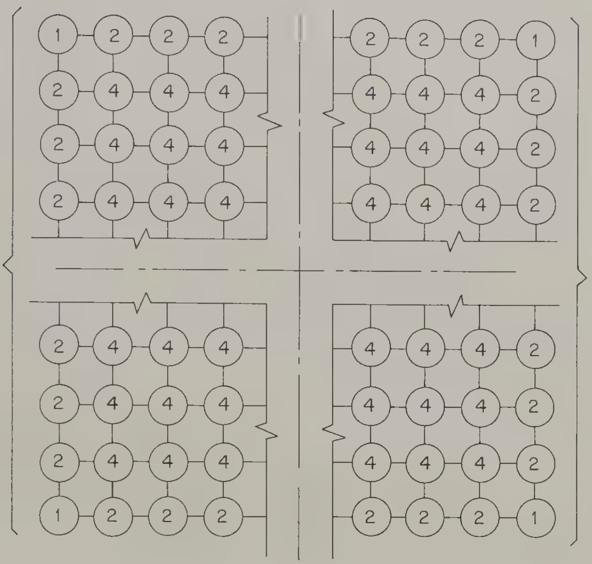
$$\frac{\partial x}{\partial u} = +1 \qquad \frac{\partial x}{\partial v} = \cos\alpha$$
$$\frac{\partial y}{\partial u} = 0 \qquad \frac{\partial y}{\partial v} = \sin\alpha \tag{11.23}$$

If, for example, a function φ were given in terms of an xy coordinate system [that is, $\varphi(x,y)$], Eqs. (11.21) and (11.22) and the relationships given by Eqs. (11.23) could be used to define the corresponding function $\varphi(u,v)$ and its partial derivatives with respect to these "new" coordinate

Table 11.3 Computational Molecules (Square Grid): Determination of Volumes (Errors of the Order of h^2)

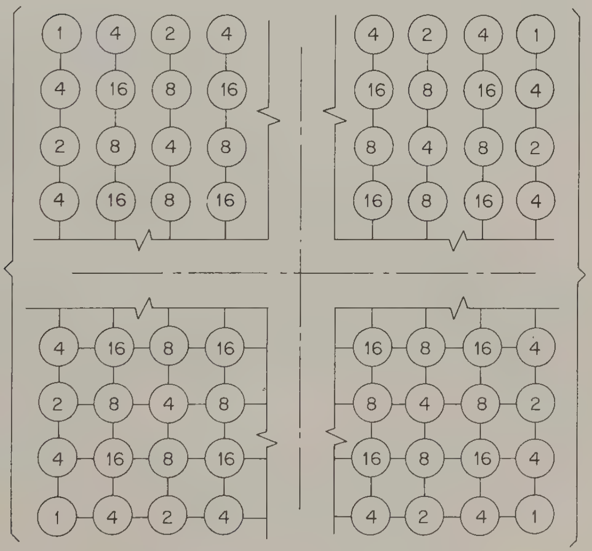
Trapezoidal rule

$$\iint_A f(x, y)\,dx\,dy \approx \frac{h^2}{4}$$



Simpson's one-third rule

$$\iint_A f(x, y)\,dx\,dy \approx \frac{h^2}{9}$$



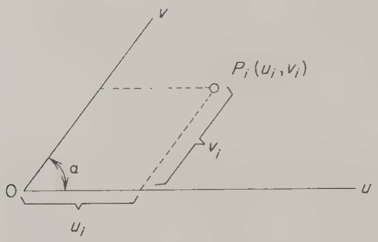


Figure 11.6

directions. The partial derivatives would be

$$\frac{\partial\varphi}{\partial u} = \frac{\partial\varphi}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial\varphi}{\partial y}\frac{\partial y}{\partial u} = \frac{\partial\varphi}{\partial x}(+1) + \frac{\partial\varphi}{\partial y}(0) = \frac{\partial\varphi}{\partial x} \tag{11.24}$$

$$\frac{\partial\varphi}{\partial v} = \frac{\partial\varphi}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial\varphi}{\partial y}\frac{\partial y}{\partial v} = \frac{\partial\varphi}{\partial x}(\cos\alpha) + \frac{\partial\varphi}{\partial y}(\sin\alpha) \tag{11.25}$$

$$\frac{\partial^2\varphi}{\partial u^2} = \frac{\partial}{\partial u}\left(\frac{\partial\varphi}{\partial u}\right) = \frac{\partial^2\varphi}{\partial x^2} \tag{11.26}$$

$$\frac{\partial^2\varphi}{\partial v^2} = \frac{\partial}{\partial v}\left(\frac{\partial\varphi}{\partial v}\right) = \frac{\partial^2\varphi}{\partial x^2}\cos^2\alpha + 2\frac{\partial^2\varphi}{\partial x\,\partial y}\sin\alpha\cos\alpha + \frac{\partial^2\varphi}{\partial y^2}\sin^2\alpha \tag{11.27}$$

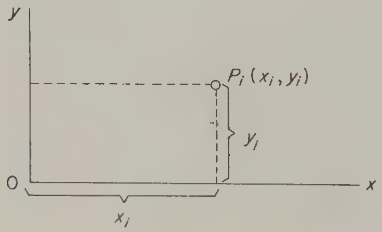
$$\frac{\partial^2\varphi}{\partial u\,\partial v} = \frac{\partial}{\partial u}\left(\frac{\partial\varphi}{\partial v}\right) = \frac{\partial^2\varphi}{\partial x^2}\cos\alpha + \frac{\partial^2\varphi}{\partial x\,\partial y}\sin\alpha \tag{11.28}$$

In matrix form, the second partial derivatives with respect to the two different coordinate systems are related as in Eqs. (11.29).

$$\begin{bmatrix} \dfrac{\partial^2\varphi}{\partial u^2} \\ \dfrac{\partial^2\varphi}{\partial v^2} \\ \dfrac{\partial^2\varphi}{\partial u\,\partial v} \end{bmatrix} = \begin{bmatrix} +1 & 0 & 0 \\ \cos^2\alpha & \sin^2\alpha & 2\sin\alpha\cos\alpha \\ \cos\alpha & 0 & \sin\alpha \end{bmatrix} \begin{bmatrix} \dfrac{\partial^2\varphi}{\partial x^2} \\ \dfrac{\partial^2\varphi}{\partial y^2} \\ \dfrac{\partial^2\varphi}{\partial x\,\partial y} \end{bmatrix} \tag{11.29}$$

The inverted system is given by Eqs. (11.30).

Figure 11.7



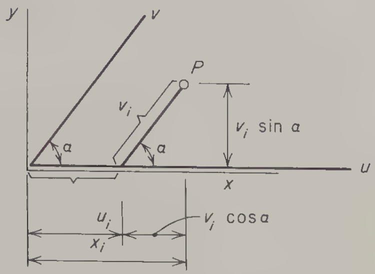


Figure 11.8

$$\begin{bmatrix} \dfrac{\partial^2\varphi}{\partial x^2} \\ \dfrac{\partial^2\varphi}{\partial y^2} \\ \dfrac{\partial^2\varphi}{\partial x\,\partial y} \end{bmatrix} = \begin{bmatrix} +1 & 0 & 0 \\ +\dfrac{\cos^2\alpha}{\sin^2\alpha} & +\dfrac{1}{\sin^2\alpha} & -\dfrac{2\cos\alpha}{\sin^2\alpha} \\ -\dfrac{\cos\alpha}{\sin\alpha} & 0 & +\dfrac{1}{\sin\alpha} \end{bmatrix} \begin{bmatrix} \dfrac{\partial^2\varphi}{\partial u^2} \\ \dfrac{\partial^2\varphi}{\partial v^2} \\ \dfrac{\partial^2\varphi}{\partial u\,\partial v} \end{bmatrix} \tag{11.30}$$

The laplacian defined in the xy system can now be expressed in terms of the "skewed" uv coordinate directions.

$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = \frac{\partial^2\varphi}{\partial u^2} + \frac{1}{\sin^2\alpha}\left(\cos^2\alpha\,\frac{\partial^2\varphi}{\partial u^2} + \frac{\partial^2\varphi}{\partial v^2} - 2\cos\alpha\,\frac{\partial^2\varphi}{\partial u\,\partial v}\right)$$

or

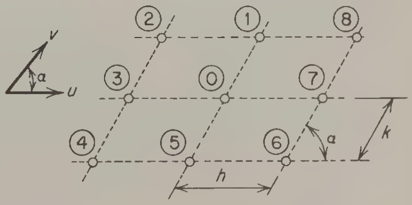
$$\nabla^2\varphi = \frac{1}{\sin^2\alpha}\left(\frac{\partial^2\varphi}{\partial u^2} - 2\cos\alpha\,\frac{\partial^2\varphi}{\partial u\,\partial v} + \frac{\partial^2\varphi}{\partial v^2}\right) \tag{11.31}$$

Note that if $\alpha = 90°$ (that is, the coordinate system is a cartesian one), Eq. (11.31) reduces to

$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial u^2} + \frac{\partial^2\varphi}{\partial v^2}$$

By using Eq. (11.31) a computational molecule whose error is of the order of h^2 can be developed. Again, for the initial definition, pivotal locations are referenced by numbers. The particular ones and the arrangement selected are shown in Fig. 11.9. In this development, at

Figure 11.9



this stage, it will not be assumed that the interval h is equal necessarily to k. Using the information contained in Table 9.27,

$$\frac{\partial^2\varphi}{\partial u^2} = \frac{1}{h^2}(\varphi_3 - 2\varphi_0 + \varphi_7) \tag{11.32}$$

$$\frac{\partial^2\varphi}{\partial u\,\partial v} = \frac{1}{4hk}(\varphi_4 - \varphi_6 - \varphi_2 + \varphi_8) \tag{11.33}$$

$$\frac{\partial^2\varphi}{\partial v^2} = \frac{1}{k^2}(\varphi_5 - 2\varphi_0 + \varphi_1) \tag{11.34}$$

Substituting these values into Eq. (11.31) yields

$$\nabla^2\varphi = \frac{1}{\sin^2\alpha}\left[\frac{1}{h^2}(\varphi_3 - 2\varphi_0 + \varphi_7) - \frac{2\cos\alpha}{4hk}(\varphi_4 - \varphi_6 - \varphi_2 + \varphi_8) + \frac{1}{k^2}(\varphi_5 - 2\varphi_0 + \varphi_1)\right] \tag{11.35}$$

If there is introduced a nondimensional parameter $\gamma = k/h$, Eq. (11.35) can be written as

$$\nabla^2\varphi = \frac{1}{2h^2\gamma^2\sin^2\alpha}\{\varphi_0[-4(1+\gamma^2)] + \varphi_1(+2) + \varphi_2(\gamma\cos\alpha) + \varphi_3(2\gamma^2) + \varphi_4(-\gamma\cos\alpha) + \varphi_5(+2) + \varphi_6(\gamma\cos\alpha) + \varphi_7(2\gamma^2) + \varphi_8(-\gamma\cos\alpha)\} \tag{11.36}$$

The computational-molecule representation is then that shown in Fig. 11.10. Assuming $h = k$, Table 11.4 summarizes "molecules" for $\nabla^2\varphi$ for angles of α equal to 30, 45, 60, and 90°.

Figure 11.10

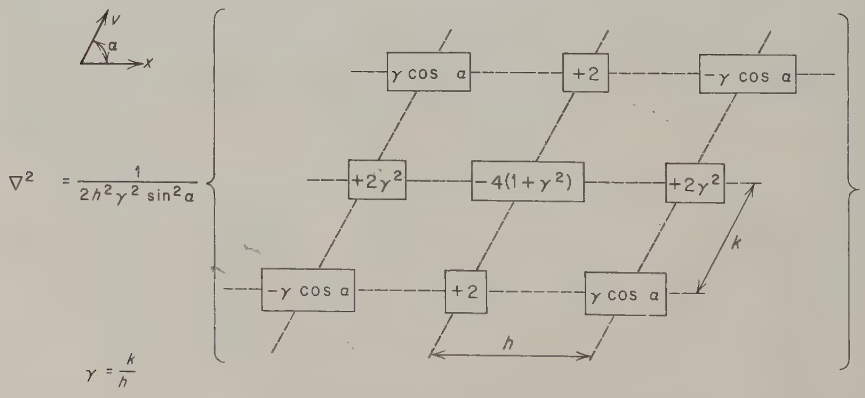
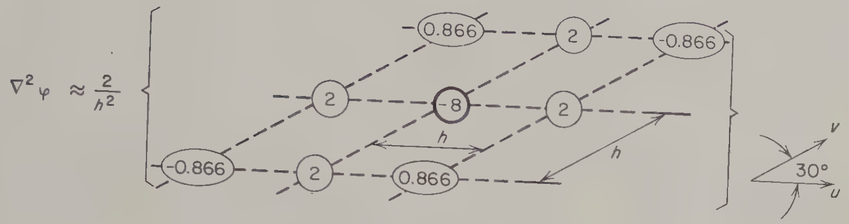
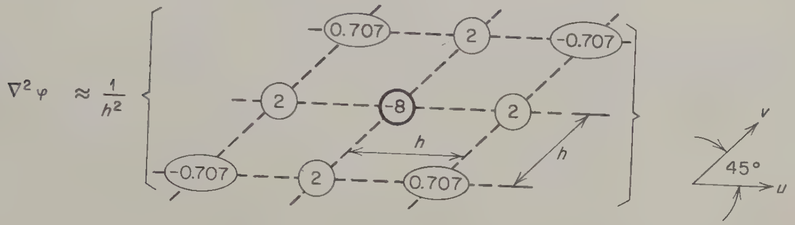


Table 11.4 Computational Molecules ["Skewed" Grid ($h = k$)]: Symmetrical Partial Finite-difference Operators for $\nabla^2\varphi$ (Errors of the Order of h^2)

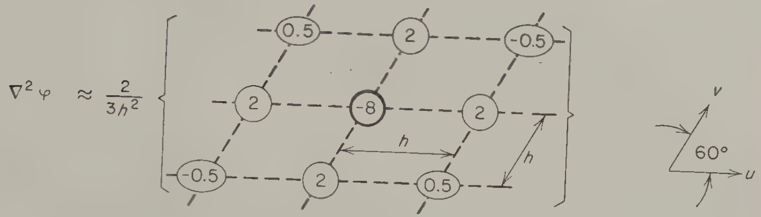
α = 30°



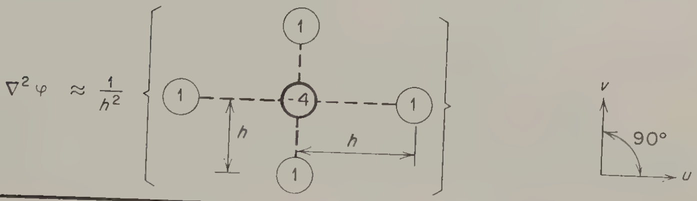
α = 45°



α = 60°



α = 90°



11.3.2 Triangular-coordinate Molecules

Another noncartesian coordinate system frequently used consists of a triangular lattice of the type shown in Fig. 11.11.

Using a procedure similar to that employed in Sec. 11.3.1, the two coordinate systems can be related, as shown in Fig. 11.12.

$$x = u + v \cos \alpha + w \cos \beta \tag{11.37}$$

$$y = v \sin \alpha + w \sin \beta \tag{11.38}$$

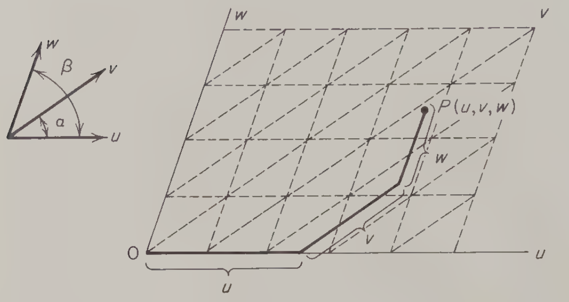


Figure 11.11

The partial derivatives of x and y with respect to u, v, and w are listed as Eqs. (11.39) and (11.40).

$$\frac{\partial x}{\partial u} = +1 \qquad \frac{\partial x}{\partial v} = \cos\alpha \qquad \frac{\partial x}{\partial w} = \cos\beta \tag{11.39}$$

$$\frac{\partial y}{\partial u} = 0 \qquad \frac{\partial y}{\partial v} = \sin\alpha \qquad \frac{\partial y}{\partial w} = \sin\beta \tag{11.40}$$

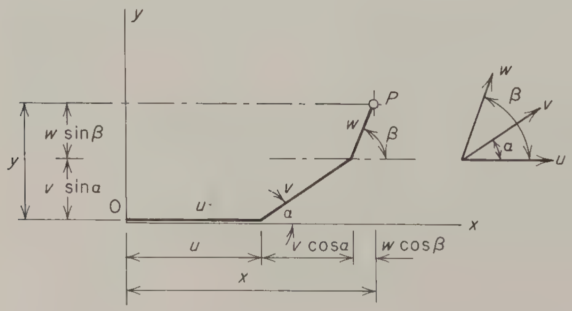
Using the same procedure as for the "skewed" case,

$$\frac{\partial^2\varphi}{\partial u^2} = \frac{\partial^2\varphi}{\partial x^2} \tag{11.41}$$

$$\frac{\partial^2\varphi}{\partial v^2} = \frac{\partial^2\varphi}{\partial x^2}\cos^2\alpha + 2\frac{\partial^2\varphi}{\partial x\,\partial y}\sin\alpha\cos\alpha + \frac{\partial^2\varphi}{\partial y^2}\sin^2\alpha \tag{11.42}$$

$$\frac{\partial^2\varphi}{\partial w^2} = \frac{\partial^2\varphi}{\partial x^2}\cos^2\beta + 2\frac{\partial^2\varphi}{\partial x\,\partial y}\sin\beta\cos\beta + \frac{\partial^2\varphi}{\partial y^2}\sin^2\beta \tag{11.43}$$

Figure 11.12



This gives for the laplacian

$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2}$$

$$= \frac{1}{2\sin\alpha\sin\beta\sin(\beta-\alpha)}\left[\frac{\partial^2\varphi}{\partial u^2}\sin 2(\beta-\alpha) - \frac{\partial^2\varphi}{\partial v^2}\sin 2\beta + \frac{\partial^2\varphi}{\partial w^2}\sin 2\alpha\right] \tag{11.44}$$

For an equilateral triangular-coordinate system, where $\alpha = 60°$ and $\beta = 120°$, Eq. (11.44) reduces to

$$\nabla^2 y = \frac{2}{3}\left(\frac{\partial^2\varphi}{\partial u^2} + \frac{\partial^2\varphi}{\partial v^2} + \frac{\partial^2\varphi}{\partial w^2}\right) \tag{11.45}$$

Computational molecules for this and the corresponding biharmonic operator are given in Table 11.5.

11.3.3 Polar Coordinates

In polar coordinates, where the point P_i is defined in terms of a distance ρ and an angle θ, the transformation equations are obtained from a consideration of Fig. 11.13.

$$x = \rho\cos\theta \tag{11.46}$$

$$y = \rho\sin\theta \tag{11.47}$$

By using the process hereinbefore described for both the "skewed" and the "triangular grid" case, it can be demonstrated that the laplacian in polar coordinates is given by

$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial\varphi}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2\varphi}{\partial\theta^2} \tag{11.48}$$

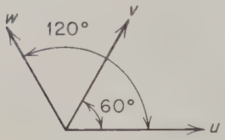
With the notation of Fig. 11.14, the computational molecule for the polar coordinates is that given by Fig. 11.15*a*, where $\gamma = k/h$. Frequently, these computational molecules are also shown in the form given in Fig. 11.15*b*.

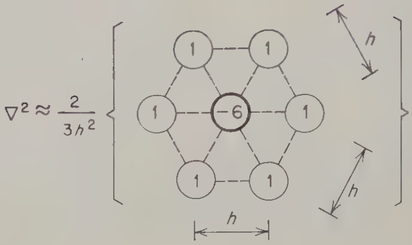
11.4 Irregular Boundaries

For most equilibrium-type problems *the value of the dependent function is specified along the boundary.* In the torsional analysis of a shaft, for example, where it is required that $\nabla^2\varphi = C$ within the interior of the region in question, it is necessary that the function φ have a certain specified constant value on the boundary.

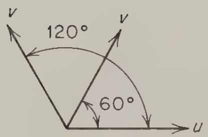
Table 11.5 Computational Molecules (Equilateral Triangular Grid): Symmetrical Partial Finite-difference Operators

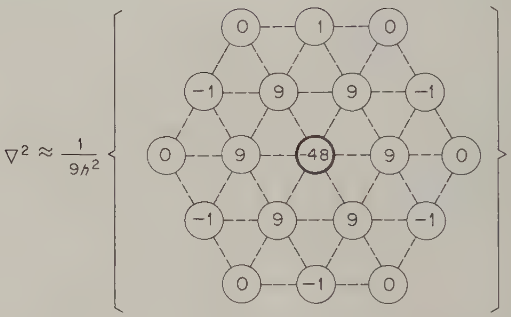
Error of the order of h^2



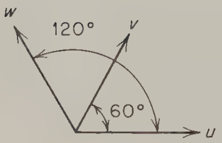


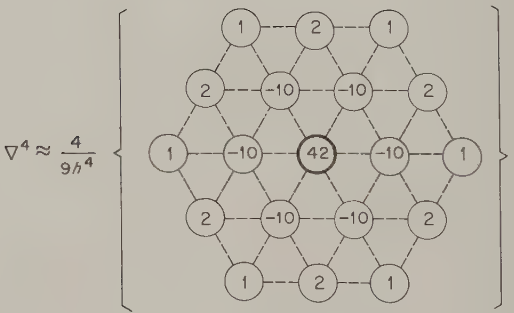
Error of the order of h^4





Error of the order of h^2





In Fig. 11.2 it was illustrated that not all boundaries necessarily coincide with pivotal locations associated with regular grids. While there are available several different methods for handling such situations, the one that will be here illustrated presumes that it is desirable to develop a computational molecule which makes use of the actual boundary

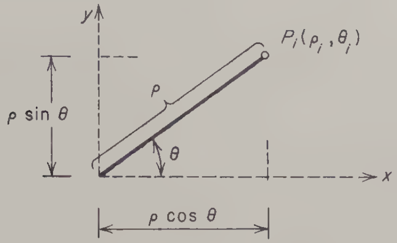


Figure 11.13

points, rather than those which fall "outside" and are associated with the continued regular grid. The cartesian-coordinate case is illustrated in Fig. 11.16. Points ⓪, ①, and ② fall within the region in question, whereas pivotal locations ③ and ④ lie outside. (The assumed boundary has been shown crosshatched.) The problem is to develop the required computational molecule for location ⓪, presuming a nonsymmetrical grid pattern defined by locations ⓪, ①, ②, ⓐ, and ⓑ. Such patterns are often referred to as *unequal-leg stars* or *unequal-leg star points.*

If it can be presumed that the function in question, $\varphi(x,y)$, can be represented to a sufficient degree of accuracy by a second-order expression in each of the coordinate directions, then it can be shown that

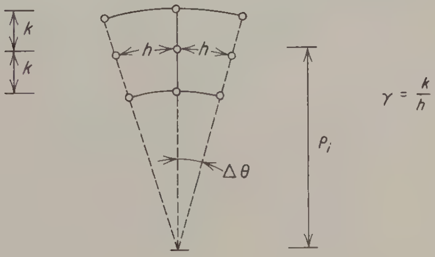
$$\frac{\partial \varphi}{\partial x} = \frac{1}{h}\left[-\frac{\alpha}{1+\alpha}\varphi_1 + \frac{1-\alpha}{\alpha}\varphi_0 + \frac{1}{\alpha(1+\alpha)}\varphi_a\right] \tag{11.49}$$

$$\frac{\partial \varphi}{\partial y} = \frac{1}{h}\left[-\frac{\beta}{1+\beta}\varphi_2 + \frac{1-\beta}{\beta}\varphi_0 + \frac{1}{\beta(1+\beta)}\varphi_b\right] \tag{11.50}$$

$$\frac{\partial^2 \varphi}{\partial x^2} = \frac{2}{h^2}\left[\frac{\varphi_1}{1+\alpha} - \frac{\varphi_0}{\alpha} + \frac{\varphi_a}{\alpha(1+\alpha)}\right] \tag{11.51}$$

$$\frac{\partial^2 \varphi}{\partial y^2} = \frac{2}{h^2}\left[\frac{\varphi_2}{1+\beta} - \frac{\varphi_0}{\beta} + \frac{\varphi_b}{\beta(1+\beta)}\right] \tag{11.52}$$

Figure 11.14



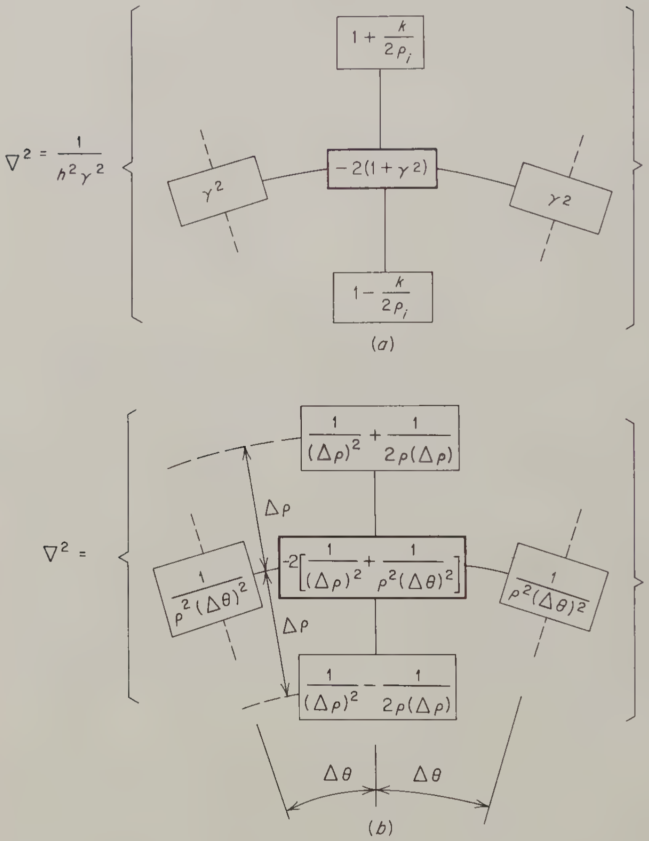
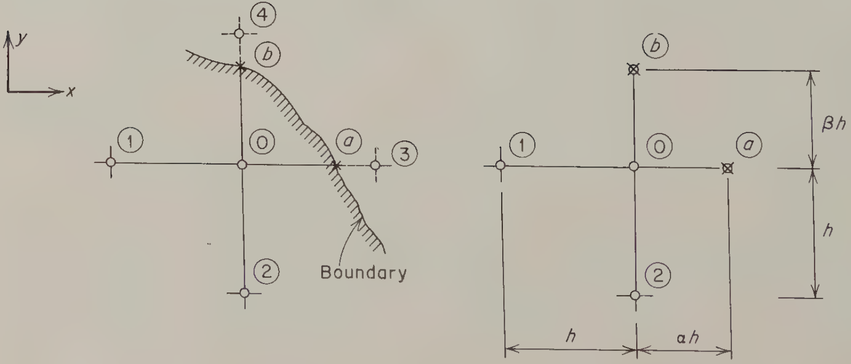


Figure 11.15

Figure 11.16



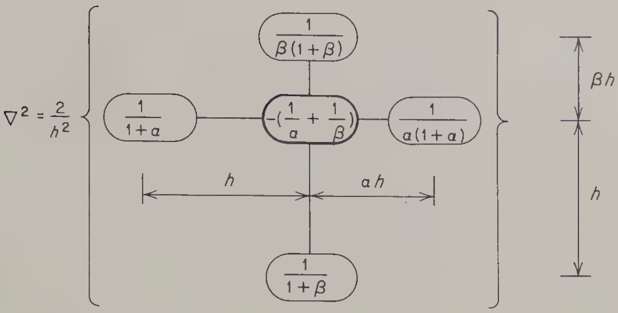


Figure 11.17

The laplacian is

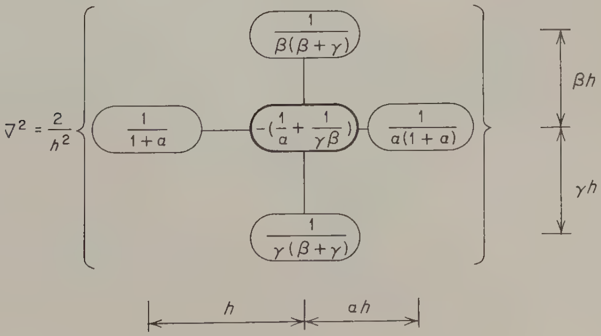
$$\nabla^2\varphi = \frac{2}{h^2}\left[\frac{\varphi_1}{1+\alpha} + \frac{\varphi_2}{1+\beta} - \frac{\varphi_0}{\alpha} - \frac{\varphi_0}{\beta} + \frac{\varphi_a}{\alpha(1+\alpha)} + \frac{\varphi_b}{\beta(1+\beta)}\right] \quad (11.53)$$

The associated computational molecule is shown in Fig. 11.17. For the case where three of the "points" of the "star" are located away from the ⓪ location at distances less than h, the computational molecule is given by Fig. 11.18.

As an example of how these molecules would be used, consider the specific case defined in Fig. 11.19. With the information contained in Fig. 11.17, the value of the laplacian at location ⓪, in terms of the function values at points ①, ②, ⓐ, and ⓑ, is that given in Eq. (11.54).

$$\nabla^2\varphi = \frac{2}{h^2}\left[-\varphi_0\left(\frac{1}{0.8} + \frac{1}{0.6}\right) + \varphi_1\frac{1}{1+0.8} + \varphi_2\frac{1}{1+0.6} + \varphi_a\frac{1}{0.8(1+0.8)} + \varphi_b\frac{1}{0.6(1+0.6)}\right]$$

Figure 11.18



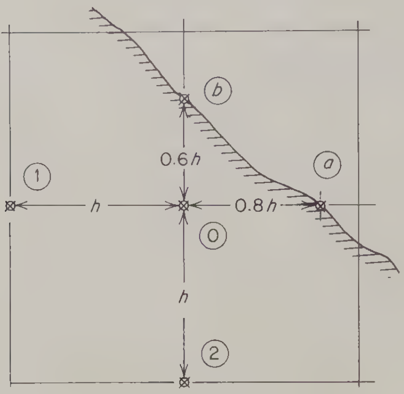


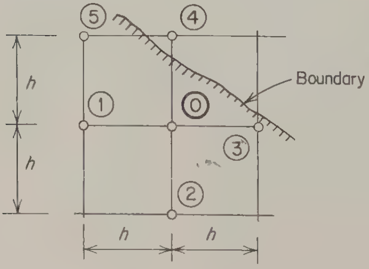
Figure 11.19

or $$\nabla^2\varphi = \frac{2}{h^2}(-2.92\varphi_0 + 0.56\varphi_1 + 0.63\varphi_2 + 0.69\varphi_a + 1.04\varphi_b) \qquad (11.54)$$

For those situations where *on the boundary the value of the normal derivative is specified,* rather than the value of the function itself (as was presumed in the preceding cases), a somewhat different approach is often used. For illustration of the method, consider the particular situation defined by Fig. 11.20. It will be assumed that the point of expansion is location ⓪. Adjacent points ①, ②, and ③ are presumed to lie within the domain in question, whereas point ④ falls without. It should be recognized that the direction of the normal to the boundary does not coincide with the assumed grid directions.

Figure 11.21 is an enlargement of a part of Fig. 11.20. If there is constructed through point ④ a perpendicular to the boundary (intersection with the boundary is presumed at point ⓐ), it is possible to define the function value at location ④ in terms of the function values at ⓪ and ⑤ and the value of the normal derivative at ⓐ.

Figure 11.20



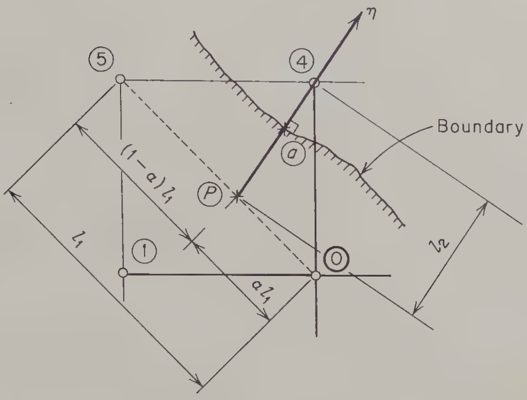


Figure 11.21

By referring to Fig. 11.21 and assuming linear interpolation, the value of the function at some location P along a line joining the pivotal points ⓪ and ⑤ can be expressed as

$$\varphi_P = \varphi_5 + (\varphi_0 - \varphi_5)(1 - \alpha) \tag{11.55}$$

or

$$\varphi_P = \alpha(\varphi_5) + (1 - \alpha)\varphi_0 \tag{11.56}$$

If linear interpolation is again assumed, the value of the normal derivative at location ⓐ is

$$\left(\frac{\partial \varphi}{\partial \eta}\right)_{\textcircled{a}} \doteq \frac{\varphi_4 - \varphi_P}{\ell_2} \tag{11.57}$$

or

$$\varphi_P = \varphi_4 - \ell_2 \left(\frac{\partial \varphi}{\partial \eta}\right)_{\textcircled{a}} \tag{11.58}$$

Eliminating φ_P from Eqs. (11.56) and (11.58) yields an expression for φ_4 in terms of φ_0, φ_5, and $(\partial\varphi/\partial\eta)_{\textcircled{a}}$.

$$\varphi_4 = (1 - \alpha)\varphi_0 + \alpha\varphi_5 + \ell_2 \left(\frac{\partial \varphi}{\partial \eta}\right)_{\textcircled{a}} \tag{11.59}$$

Equation (11.59) can be used directly in the establishment of modified finite-difference expressions. For example, the laplacian

$$\nabla^2\varphi = \frac{1}{h^2}(\varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 - 4\varphi_0) \tag{11.60}$$

becomes

$$\nabla^2\varphi = \frac{1}{h^2}\left\{\varphi_1 + \varphi_2 + \varphi_3 + \alpha\varphi_5 + \ell_2\left(\frac{\partial \varphi}{\partial \eta}\right)_{\textcircled{a}} - [4 - (1 - \alpha)\varphi_0]\right\} \tag{11.61}$$

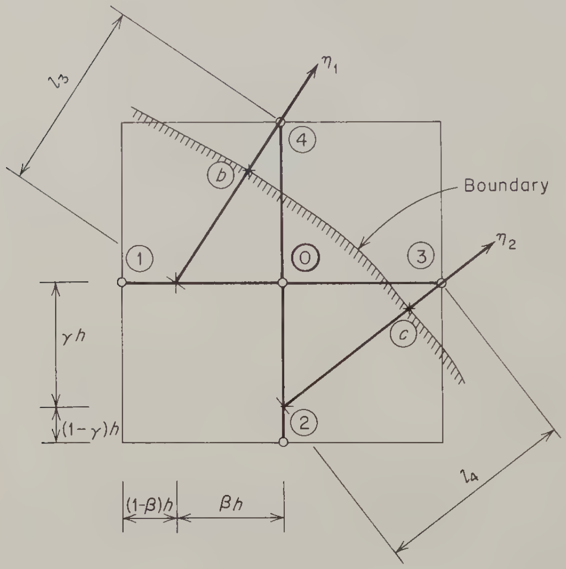


Figure 11.22

When two points fall outside the domain, a situation of the type defined in Fig. 11.22 must be considered. Using the same procedure as in the previous case yields the following:

$$\varphi_3 = (1-\gamma)\varphi_0 + \gamma\varphi_2 + \ell_4 \left(\frac{\partial\varphi}{\partial\eta_2}\right)_{\textcircled{c}} \tag{11.62}$$

$$\varphi_4 = (1-\beta)\varphi_0 + \beta\varphi_1 + \ell_3 \left(\frac{\partial\varphi}{\partial\eta_1}\right)_{\textcircled{b}} \tag{11.63}$$

The laplacian then becomes

$$\nabla^2\varphi = \frac{1}{h^2}\left\{(1+\beta)\varphi_1 + (1+\gamma)\varphi_2 + \ell_3\left(\frac{\partial\varphi}{\partial\eta_1}\right)_{\textcircled{b}} + \ell_4\left(\frac{\partial\varphi}{\partial\eta_2}\right)_{\textcircled{c}} \right.$$
$$\left. - [4-(1-\beta)-(1-\gamma)]\varphi_0\right\} \tag{11.64}$$

When the point of expansion of a "star" is located on the boundary, and it is required that the normal derivative at that point take on a particular value, a still different situation must be examined (see Fig. 11.23). For this case, the laplacian becomes

$$\nabla^2\varphi = \frac{1}{h^2}\left\{2\varphi_2 + \delta\varphi_6 + \epsilon\varphi_7 + \ell_7\left(\frac{\partial\varphi}{\partial\eta_3}\right)_{\textcircled{d}} + \ell_8\left(\frac{\partial\varphi}{\partial\eta_4}\right)_{\textcircled{e}} + 2h\left(\frac{\partial\varphi}{\partial\eta_5}\right)_{\textcircled{0}} \right.$$
$$\left. - [4-(1-\delta)-(1-\epsilon)]\varphi_0\right\} \tag{11.65}$$

It should be recognized that when compared with points remote from the boundaries, Eqs. (11.61), (11.64), and (11.65) are less accurate. While it is possible to develop other, more accurate methods for handling the normal-derivative boundary-value problem, these other methods will not be considered in this introductory text.

11.5 Elliptic Partial Differential Equations: Boundary-value Problems

Example 11.1 As a first example of the numerical solution of two-dimensional elliptic boundary-value-type partial differential equations, consider the problem defined in Fig. 11.24. The region of interest is square and has a side dimension of a. The specified "edge loadings" are:

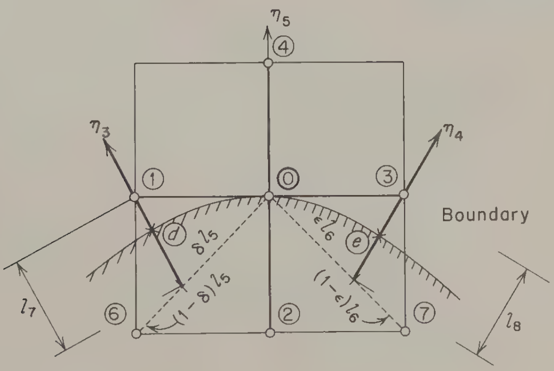
$$\begin{aligned}
&\text{Along } x = 0: && \varphi = 1.0 - \frac{y}{a} \\
&\text{Along } x = a: && \varphi = 0 \\
&\text{Along } y = 0: && \varphi = 1.0 - \frac{x}{a} \\
&\text{Along } y = a: && \varphi = 0
\end{aligned} \tag{11.66}$$

Within the given boundaries, it is desired to determine the distribution of φ such that

$$\nabla^2\varphi = 0 \tag{11.67}$$

It will first be assumed that a three-equal-part division of each of the coordinate dimensions (that is, $h = a/3$) will result in a sufficiently accurate solution. The chosen grid system is shown in Fig. 11.25.

Figure 11.23



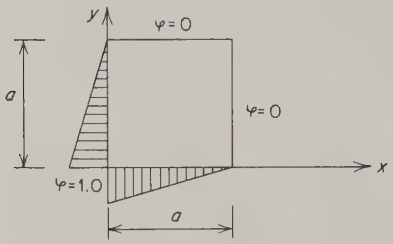
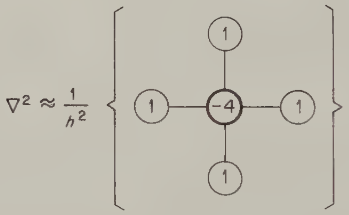


Figure 11.24

Four unknown values of φ, at pivotal points ①, ②, ③, and ④, are presumed. At each of these locations, it is required that the laplacian (in finite-difference form) equal zero.

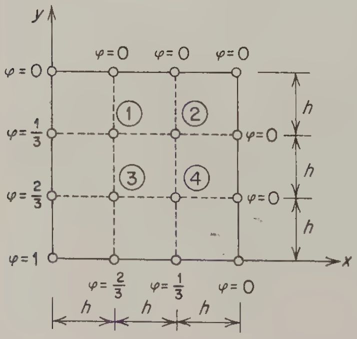
From Table 11.1, the computational molecule associated with ∇^2 is



(11.68)

Since Eq. (11.67) requires that $\nabla^2\varphi = 0$, it is sufficient to require at each of the interior points of expansion that the bracketed term in Eq. (11.68) equal zero. Consider for illustration expansion about location ①, as shown in Fig. 11.26. The unknown values of the functions φ at each

Figure 11.25



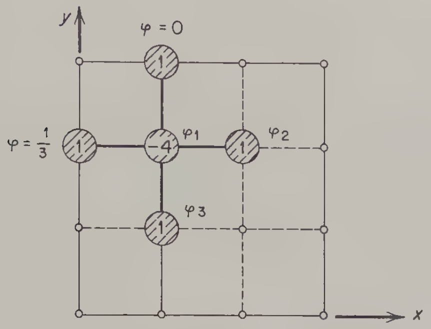


Figure 11.26

location are also noted in the figure. The corresponding equation is

$$(+1)(\tfrac{1}{3}) + (+1)(\varphi_3) + (+1)(\varphi_2) + (+1)(0) + (-4)(\varphi_1) = 0 \quad (11.69)$$

Similarly, equations are written for each of the other interior points. The resulting set of equations is

$$\begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 1 & 0 & -4 & 1 \\ 0 & 1 & 1 & -4 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \end{bmatrix} = \begin{bmatrix} -\tfrac{1}{3} \\ 0 \\ -\tfrac{4}{3} \\ -\tfrac{1}{3} \end{bmatrix} \quad (11.70)$$

The solution, which can be obtained by any of the methods defined in Chap. 5, is

$$\varphi_1 = 0.22 \qquad \varphi_2 = 0.11 \qquad \varphi_3 = 0.44 \qquad \varphi_4 = 0.22 \quad (11.71)$$

This solution is illustrated in Fig. 11.27.

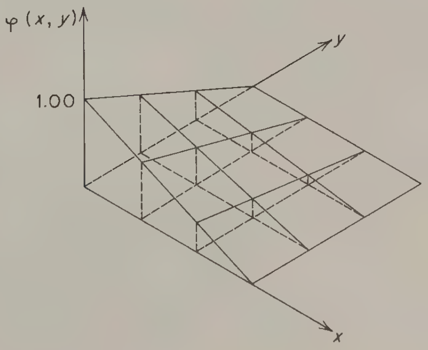


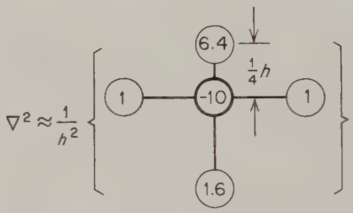
Figure 11.27

Example 11.2 As a second example, the Poissonian equation

$$\nabla^2\varphi + 2 = 0$$

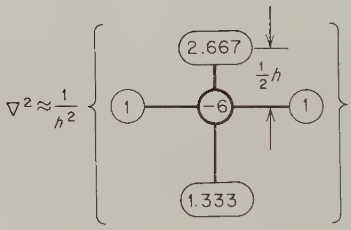
is prescribed for the interior of the region defined in Fig. 11.28. On the boundary it is required that $\varphi = 0$. (The selected grid system has been indicated on the figure.) It should be noted that points ①, ②, and ③ require the development of unequal-leg stars for the Laplace operator. With the information contained in Fig. 11.17, these are found to be:

Location ①:



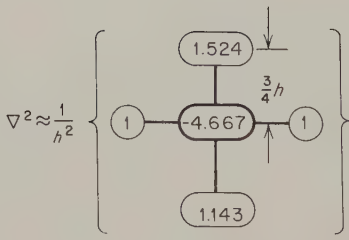
(11.72)

Location ②:



(11.73)

Location ③:



(11.74)

The equations governing the solution are listed as Eqs. (11.75).

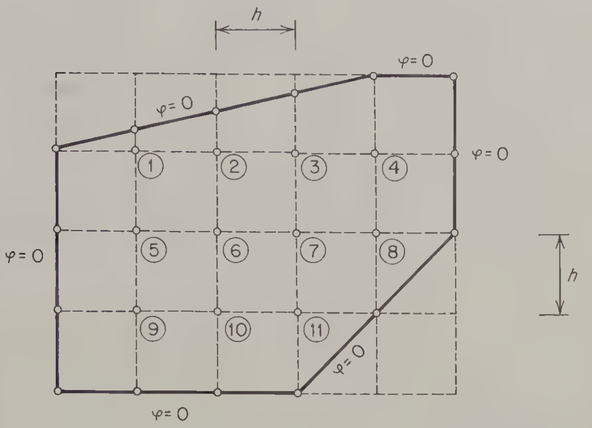


Figure 11.28

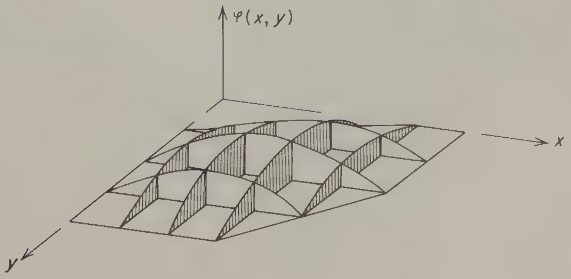
$$\begin{bmatrix} -10 & 1 & 0 & 0 & 1.6 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -6 & 1 & 0 & 0 & 1.333 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4.667 & 1 & 0 & 0 & 1.143 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -4 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -4 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & -4 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & -4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \\ \varphi_8 \\ \varphi_9 \\ \varphi_{10} \\ \varphi_{11} \end{bmatrix} = \begin{bmatrix} -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \\ -2h^2 \end{bmatrix} \tag{11.75}$$

The solution is

$$\begin{aligned} \varphi_1 &= 0.551h^2 & \varphi_5 &= 1.488h^2 & \varphi_9 &= 1.300h^2 \\ \varphi_2 &= 1.131h^2 & \varphi_6 &= 2.101h^2 & \varphi_{10} &= 1.712h^2 \\ \varphi_3 &= 1.433h^2 & \varphi_7 &= 2.074h^2 & \varphi_{11} &= 1.446h^2 \\ \varphi_4 &= 1.187h^2 & \varphi_8 &= 1.315h^2 \end{aligned} \tag{11.76}$$

which is illustrated in Fig. 11.29.

Figure 11.29



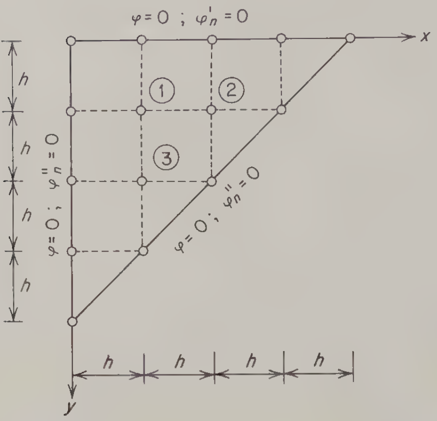


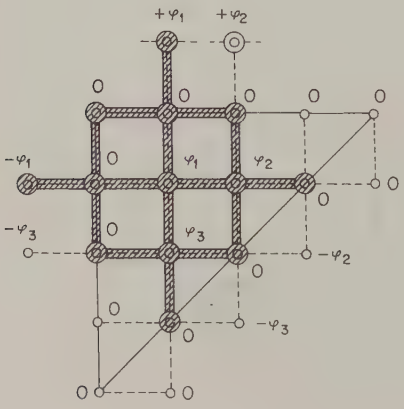
Figure 11.30

Example 11.3 As an example of a higher-order boundary-value-type partial differential equation problem, consider the solution of the following biharmonic equation:

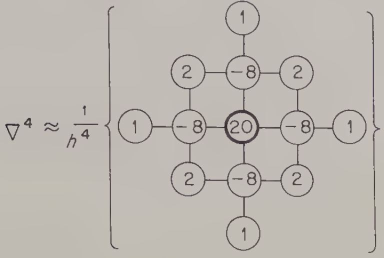
$$\nabla^4\varphi = \frac{\partial^4\varphi}{\partial x^4} + 2\,\frac{\partial^4\varphi}{\partial x^2\,\partial y^2} + \frac{\partial^4\varphi}{\partial y^4} = K \tag{11.77}$$

The domain in question is shown in Fig. 11.30, as are the prescribed boundary conditions and the selected grid system. Three unknown pivotal values have been assumed: φ_1, φ_2, and φ_3. To facilitate the handling of the specified boundary conditions, imaginary pivotal locations outside the domain will be presumed. Figure 11.31 shows these imaginary locations and their corresponding φ values—related to the

Figure 11.31



interior points—as well as the general form of the computational molecule for ∇^4 placed "over" location ① for evaluation. From Table 11.1, the computational molecule to be used for ∇^4 is



(11.78)

The controlling equations are therefore

$$\begin{bmatrix} 20 & -8 & -8 \\ -8 & 19 & 1 \\ -8 & 1 & 17 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{bmatrix} = \begin{bmatrix} h^4K \\ h^4K \\ h^4K \end{bmatrix} \tag{11.79}$$

It is to be noted that the special equations for handling normal-derivative boundary conditions that were derived in Sec. 11.4 were not used in this example. This was due to the coincidence of the assumed grid pattern and the normals to the boundaries, and defined zero values of the function on these boundaries.

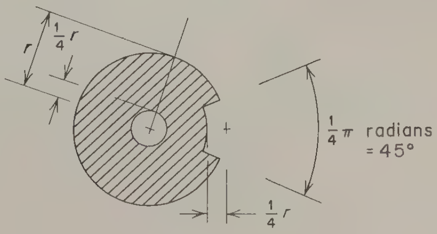
Example 11.4 To illustrate the use of the previously developed polar-coordinate computational molecule (Fig. 11.15), consider the problem of finding the distribution of a function φ over a cross section of a keyed, hollow, circular shaft, when φ is governed by the equation

$$\nabla^2\varphi = K \tag{11.80}$$

It is required that $\varphi = 0$ on the boundary. The specific case to be examined is illustrated in Fig. 11.32. It is to be recognized that symmetry exists, and therefore only one-half of the cross section need be considered in any numerical computations.

The concentric-circle-and-radial-line polar-grid pattern shown in Fig.

Figure 11.32



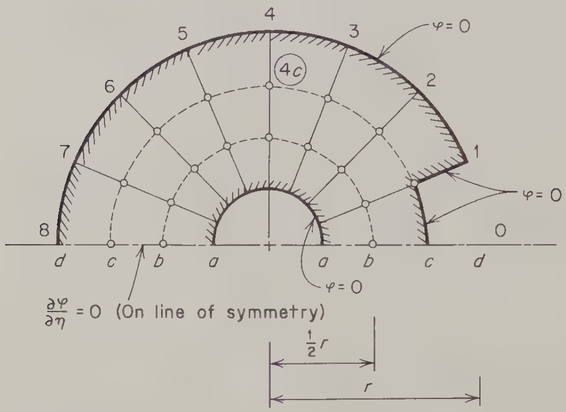
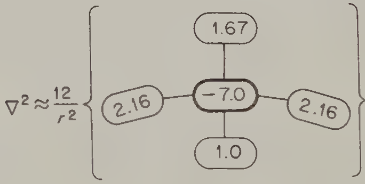

4
5
3
6
2
7
1
8
0
4c
φ = 0
φ = 0
φ = 0
d
c
b
a
a
b
c
d
$\frac{\partial\varphi}{\partial\eta}$ = 0 (On line of symmetry)
$\frac{1}{2}r$
r

Figure 11.33

11.33 will be used. For reference purposes the rings have been labeled with lowercase letters; whereas the radial lines have been numbered. Pivotal location (4c) has been noted on the sketch. The entire 360° circle has been subdivided into 16 equal parts ($\Delta\theta = \pi/8$ rad). For ring b, the ρ distance is $\frac{1}{2}r$. For ring c, the radius is $\frac{3}{4}r$.

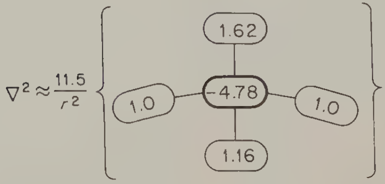
With the information shown in Fig. 11.15, the computational molecules associated with rings b and c are shown as Eqs. (11.81) and (11.82).

Ring b:



(11.81)

Ring c:



(11.82)

At location (4c), the expanded equation is

$$-4.78\varphi_{4c} + 1.00\varphi_{3c} + 1.00\varphi_{5c} + 1.16\varphi_{4b} = \frac{Kr^2}{11.5} \tag{11.83}$$

The entire set of governing equations is given as Eqs. (11.84).

$$
\begin{bmatrix}
-7.0 & 4.32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2.16 & -7.0 & 2.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.16 & -7.0 & 1.67 & 2.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.16 & -4.78 & 0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2.16 & 0 & -7.0 & 1.67 & 2.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.0 & 1.16 & -4.78 & 0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2.16 & 0 & -7.0 & 1.67 & 2.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.0 & 1.16 & -4.78 & 0 & 1.0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2.16 & 0 & -7.0 & 1.67 & 2.16 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & 1.16 & -4.78 & 0 & 1.0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.16 & 0 & -7.0 & 1.67 & 2.16 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & 1.16 & -4.78 & 0 & 1.0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.16 & 0 & -7.0 & 1.67 & 2.16 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.0 & 1.16 & -4.78 & 0 & 1.0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4.32 & 0 & -7.0 & 1.67 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.0 & 1.16 & -4.78
\end{bmatrix}
\begin{bmatrix}
\varphi_{0b} \\ \varphi_{1b} \\ \varphi_{2b} \\ \varphi_{2c} \\ \varphi_{3b} \\ \varphi_{3c} \\ \varphi_{4b} \\ \varphi_{4c} \\ \varphi_{5b} \\ \varphi_{5c} \\ \varphi_{6b} \\ \varphi_{6c} \\ \varphi_{7b} \\ \varphi_{7c} \\ \varphi_{8b} \\ \varphi_{8c}
\end{bmatrix}
=
\begin{bmatrix}
0.083Kr^2 \\ 0.083Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2 \\ 0.083Kr^2 \\ 0.087Kr^2
\end{bmatrix}
\qquad (11.84)
$$

The solution is

$$\begin{array}{lll}
\varphi_{0b} = -0.0362Kr^2 & \varphi_{3b} = -0.0612Kr^2 & \\
 & \varphi_{5b} = -0.0669Kr^2 & \varphi_{7b} = -0.0679Kr^2 \\
\varphi_{1b} = -0.0394Kr^2 & \varphi_{3c} = -0.0540Kr^2 & \\
 & \varphi_{5c} = -0.0589Kr^2 & \varphi_{7c} = -0.0596Kr^2 \\
\varphi_{2b} = -0.0530Kr^2 & \varphi_{4b} = -0.0655Kr^2 & \\
 & \varphi_{6b} = -0.0676Kr^2 & \varphi_{8b} = -0.0680Kr^2 \\
\varphi_{2c} = -0.0424Kr^2 & \varphi_{4c} = -0.0576Kr^2 & \\
 & \varphi_{6c} = -0.0594Kr^2 & \varphi_{8c} = -0.0596Kr^2
\end{array} \tag{11.85}$$

Example 11.5 The buckling of a plate is a boundary-value eigenvalue-type problem that can be conveniently solved by using the finite-difference methods heretofore developed. For illustration, consider the particular case shown in Fig. 11.34. The plate, which is square, is presumed to be of uniform thickness. It is "fixed"—that is, $\partial w/\partial y = 0$ and $w = 0$—along one of its edges, but "simply supported"—$\partial^2 w/\partial x^2 = 0$ and $w = 0$—along the other three edges. A force per unit length of N_x is assumed along two opposite edges, as shown. The problem is to determine the critical value(s) of N_x such that the plate can deflect (buckle) out of its original plane.

The governing fourth-order partial differential equation is

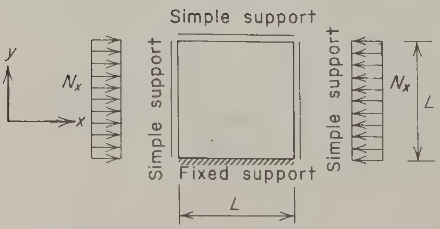
$$\nabla^4 w + \lambda \frac{\partial^2 w}{\partial x^2} = 0 \tag{11.86}$$

where w = lateral deflection of plate

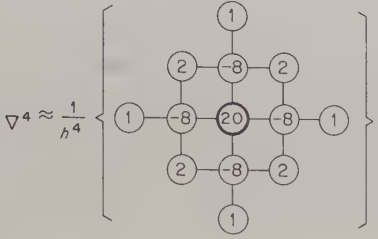
$$\lambda = \frac{N_x}{D} \tag{11.87}$$

D = plate bending stiffness

Figure 11.34

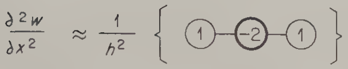


From Table 11.1 the necessary computational molecules are



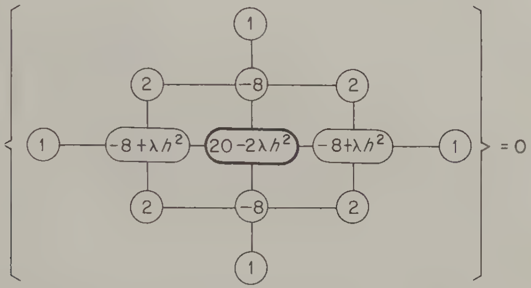
(11.88)

and



(11.89)

Equation (11.86) can therefore be expressed in finite-difference computational-molecule form.



(11.90)

If it is assumed that a sufficiently accurate solution can be obtained by dividing each of the sides of the plate into four equal parts, the grid system to be investigated is that shown in Fig. 11.35. It should be noted that imaginary pivotal points outside the plate also have been shown. By using the specified boundary conditions, these are related to the unknown interior values.

Applying Eq. (11.90) at each of the interior pivotal locations gives the set of equations (11.91) (on the following page):

$$\begin{bmatrix} 18-2\lambda h^2 & -8+\lambda h^2 & 1 & -8 & 2 & 0 & 1 & 0 & 0 \\ -8+\lambda h^2 & 19-2\lambda h^2 & -8-\lambda h^2 & 2 & -8 & 2 & 0 & 1 & 0 \\ 1 & -8-\lambda h^2 & 18-2\lambda h^2 & 0 & 2 & -8 & 0 & 0 & 1 \\ -8 & 2 & 0 & 19-2\lambda h^2 & -8+\lambda h^2 & 1 & -8 & 2 & 0 \\ 2 & -8 & 2 & -8+\lambda h^2 & 20-2\lambda h^2 & -8+\lambda h^2 & 2 & -8 & 2 \\ 0 & 2 & -8 & 1 & -8+\lambda h^2 & 19-2\lambda h^2 & 0 & 2 & -8 \\ 1 & 0 & 0 & -8 & 2 & 0 & 20-2\lambda h^2 & -8+\lambda h^2 & 1 \\ 0 & 1 & 0 & 2 & -8 & 2 & -8+\lambda h^2 & 21-2\lambda h^2 & -8+\lambda h^2 \\ 0 & 0 & 1 & 0 & 2 & -8 & 1 & -8+\lambda h^2 & 20-2\lambda h^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{11.91}$$

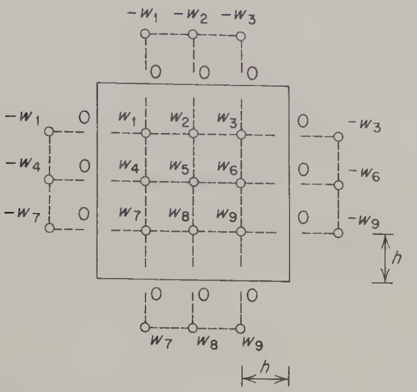


Figure 11.35

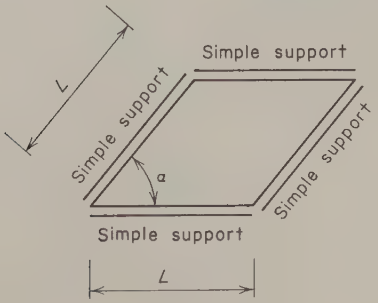
For this set, the first five eigenvalues are:

$$\begin{aligned} \lambda h^2 &= 3.032 \quad \text{or} \quad & \lambda L^2 &= 48.512 \\ &= 3.563 & &= 57.008 \\ &= 4.820 & &= 77.120 \\ &= 8.504 & &= 136.064 \\ &= 8.880 & &= 142.080 \end{aligned} \tag{11.92}$$

Example 11.6 As a second eigenvalue example, consider the determination of the natural frequency of vibration of a skewed, simply supported membrane. The dimensions of the membrane to be examined are given in Fig. 11.36. The governing partial differential equation is

$$\nabla^2 w + \lambda w = 0 \tag{11.93}$$

Figure 11.36



where w = lateral deflection of membrane

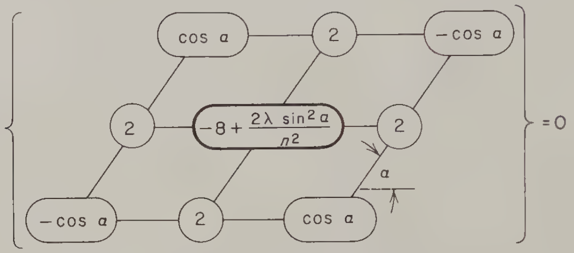
$\lambda = \dfrac{mL^2}{S}\,\omega^2$

m = mass per unit of area

S = tensile force per unit of length of membrane

ω = natural frequency

The "composite" finite-difference computational molecule associated with Eq. (11.93) is



(11.94)

$n = L/h$ = number of equal-length segments into which each of the sides is divided.

First, assume that a sufficiently accurate solution can be obtained by dividing each of the sides into two equal parts; that is, $n = 2$. This case is illustrated in Fig. 11.37. Since it is specified that $w = 0$ along the boundaries of the membrane, the application of Eq. (11.94) to the case defined by Fig. 11.37 (about location ①) yields

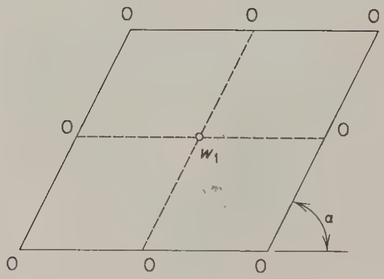
$$\left(-8+\frac{2\lambda\sin^2\alpha}{2^2}\right)w_1 = 0 \tag{11.95}$$

or

$$\lambda = \frac{16}{\sin^2\alpha} \tag{11.96}$$

Secondly, had n been selected equal to 3, the situation would be that

Figure 11.37



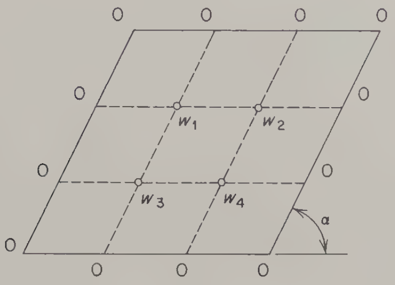


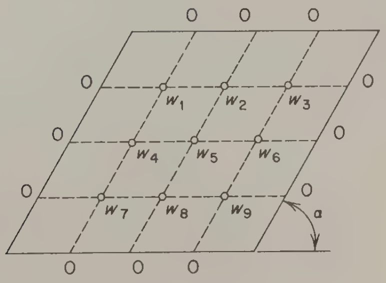
Figure 11.38

illustrated in Fig. 11.38. Applying Eq. (11.94) to each of the indicated pivotal locations results in the following set of equations:

$$\begin{bmatrix} -8+\dfrac{2\lambda\sin^2\alpha}{9} & 2 & 2 & \cos\alpha \\ 2 & -8+\dfrac{2\lambda\sin^2\alpha}{9} & -\cos\alpha & 2 \\ 2 & -\cos\alpha & -8+\dfrac{2\lambda\sin^2\alpha}{9} & 2 \\ \cos\alpha & 2 & 2 & -8+\dfrac{2\lambda\sin^2\alpha}{9} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (11.97)$$

Thirdly, Fig. 11.39 is for that case where $n = 4$. The controlling equations are those listed as Eqs. (11.98).

Figure 11.39



$$\begin{bmatrix} -8+\frac{\lambda\sin^2\alpha}{8} & 2 & 0 & 2 & \cos\alpha & 0 & 0 & 0 & 0 \\ 2 & -8+\frac{\lambda\sin^2\alpha}{8} & 2 & -\cos\alpha & 2 & \cos\alpha & 0 & 0 & 0 \\ 0 & 2 & -8+\frac{\lambda\sin^2\alpha}{8} & 0 & -\cos\alpha & 2 & 0 & 0 & 0 \\ 2 & -\cos\alpha & 0 & -8+\frac{\lambda\sin^2\alpha}{8} & 2 & 0 & 2 & \cos\alpha & 0 \\ \cos\alpha & 2 & -\cos\alpha & 2 & -8+\frac{\lambda\sin^2\alpha}{8} & 2 & -\cos\alpha & 2 & \cos\alpha \\ 0 & \cos\alpha & 2 & 0 & 2 & -8+\frac{\lambda\sin^2\alpha}{8} & 0 & -\cos\alpha & 2 \\ 0 & 0 & 0 & 2 & -\cos\alpha & 0 & -8+\frac{\lambda\sin^2\alpha}{8} & 2 & 0 \\ 0 & 0 & 0 & \cos\alpha & 2 & -\cos\alpha & 2 & -8+\frac{\lambda\sin^2\alpha}{8} & 2 \\ 0 & 0 & 0 & 0 & \cos\alpha & 2 & 0 & 2 & -8+\frac{\lambda\sin^2\alpha}{8} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{11.98}$$

For the particular two cases where $\alpha = 60°$ and $\alpha = 90°$, the smallest eigenvalues [obtained from Eqs. (11.96) to (11.98)] are given in Table 11.6. Also listed are λ values obtained by extrapolation to that limiting case where $1/n^2 = 0$.

Table 11.6

Number of Divisions		$\alpha = 60°$	$\alpha = 90°$
$n = 2$	Eq. (11.96)	$\lambda = 21.33$	$\lambda = 16.00$
$n = 3$	Eqs. (11.97)	$\lambda = 23.82$	$\lambda = 18.00$
$n = 4$	Eqs. (11.98)	$\lambda = 24.52$	$\lambda = 18.75$
Extrapolation to the case where $n = \infty$		$\lambda = 25.29$	$\lambda = 19.74$

Extrapolation was carried out by using the Lagrange formula (9.73) applied to the case illustrated in Fig. 11.40. It is assumed that there are three known values: x_1,λ_1; x_2,λ_2; and x_3,λ_3. The desired equation is

$$\lambda_0 = \frac{(x_0 - x_2)(x_0 - x_3)}{(x_1 - x_2)(x_1 - x_3)}\lambda_1 + \frac{(x_0 - x_1)(x_0 - x_3)}{(x_2 - x_1)(x_2 - x_3)}\lambda_2 + \frac{(x_0 - x_1)(x_0 - x_2)}{(x_3 - x_1)(x_3 - x_2)}\lambda_3 \qquad (11.99)$$

or

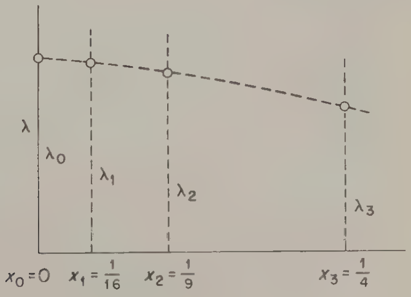
$$\lambda_0 = 3.0476\lambda_1 - 2.3143\lambda_2 + 0.2667\lambda_3 \qquad (11.100)$$

11.6 *The Use of Relaxation Methods in the Solution of Boundary-value Problems*

11.6.1 Unit Relaxation

As pointed out in Sec. 11.5, since equilibrium-type problems—solved by finite-difference methods—result in the solution of simultaneous

Figure 11.40



linear algebraic equations, any of the methods defined in Chap. 5 can be used. Relaxation methods (Sec. 5.5) are often encountered.

Example 11.1a Consider again, as a first demonstration of the method in question, Example 11.1. In relaxation form, the equations to be solved [Eqs. (11.70)] are

$$\begin{aligned} -4\varphi_1 + \varphi_2 + \varphi_3 \qquad\quad + \tfrac{1}{3} &= R_1 \\ \varphi_1 - 4\varphi_2 \qquad + \varphi_4 \qquad &= R_2 \\ \varphi_1 \qquad - 4\varphi_3 + \varphi_4 + \tfrac{4}{3} &= R_3 \\ \varphi_2 + \varphi_3 - 4\varphi_4 + \tfrac{1}{3} &= R_4 \end{aligned} \tag{11.101}$$

The unit-relaxation table is listed in Table 11.7. The relaxation solution is given in Table 11.8.

Table 11.7

$\Delta\varphi_1$	$\Delta\varphi_2$	$\Delta\varphi_3$	$\Delta\varphi_4$	ΔR_1	ΔR_2	ΔR_3	ΔR_4
1	0	0	0	$\underline{-4}$	1	1	0
0	1	0	0	1	$\underline{-4}$	0	1
0	0	1	0	1	0	$\underline{-4}$	1
0	0	0	1	0	1	1	$\underline{-4}$

Table 11.8

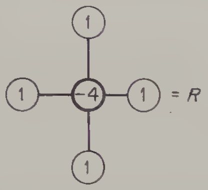
Line	φ_1	φ_2	φ_3	φ_4	R_1	R_2	R_3	R_4	Note
1	0	0	0	0	0.33	0	1.32	0.33	
2	0	0	0	0	33	0	132	33	×100
3			33		66	0	0	66	↓
4	17				−2	17	17	66	
5				17	−2	34	34	−2	
6		8			6	2	34	6	
7			8		14	2	2	14	
8	4				−2	6	6	14	
9				4	−2	10	10	−2	
10		3			1	−2	10	1	
11			3		4	−2	−2	4	
12	1				0	−1	−1	4	
13				1	0	0	0	0	
	22	11	44	22					×100

The solution is therefore

$$\varphi_1 = 0.22 \qquad \varphi_2 = 0.11 \qquad \varphi_3 = 0.44 \qquad \varphi_4 = 0.22 \tag{11.102}$$

An alternative procedure for carrying out this relaxation is frequently used for boundary-value-type partial-differential-equation problems. Actually, the relaxation procedure is the same, but the "bookkeeping" is handled in a somewhat different fashion.

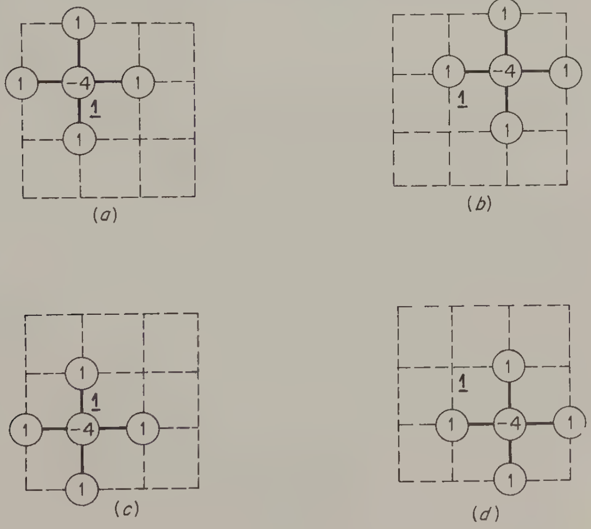
It is to be recognized at the outset that the unit-relaxation table listed as Table 11.7 can be deduced directly from a consideration of the finite-difference computational-molecule representation of the Laplace operator.



(11.103)

Consider, for example, the influence of a unit change in the value of φ_1 on the definition of the "equilibrium equations" for the four placements (see Fig. 11.41) of the "molecule" given as Eq. (11.103). For each of the first three placements, i.e., cases *a*, *b*, and *c*, the unit change in φ at location ① results in a change in the value of the laplacian, and therefore in the value of the residual. For case *d*, however, no influence is realized.

Figure 11.41 *(a)* Eq. (1); *(b)* Eq. (2); *(c)* Eq. (3); *(d)* Eq. (4).



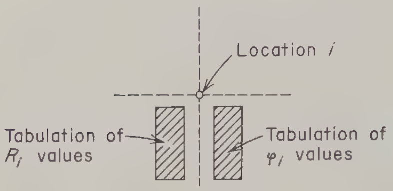


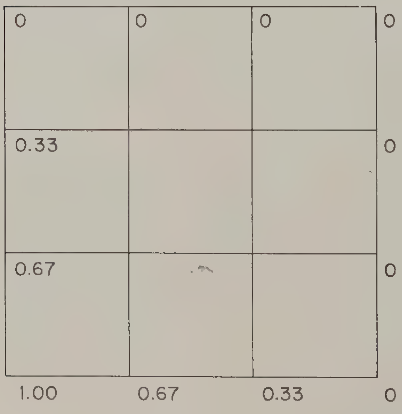
Figure 11.42

It should further be recognized from a consideration of the molecule for the laplacian [Eq. (11.103)] that a unit change in the φ value for location i results in a change in R_i of -4, and a change of $+1$ in each of the adjacent R's. This, of course, does not hold for the boundaries, where normally φ's are specified, and are therefore not subject to relaxation.

Since the influence of altering any one variable is restricted to a small region, it is comparatively easy to keep a "running account" of values of φ and R directly on a grid which represents the domain in question. For example, the arrangement shown in Fig. 11.42 is frequently used.

Example 11.1b For the example problem just considered, the "known" or given φ values are those listed in Fig. 11.43. By multiplying each of these φ values by 100 and selecting as "initial guesses" values (for 100φ) at the four unknown interior pivotal locations, possible starting values for relaxation are those given in Fig. 11.44. Note that the largest residuals (-57) occur where 100φ was assumed to be 40. A first relaxation at either one (or both) of these locations therefore should be undertaken. Since a change of $\Delta\varphi_i$ will alter R_i by a factor of $-4(\Delta\varphi_i)$, an addition of -14 to the $+40$ will reduce the residual

Figure 11.43



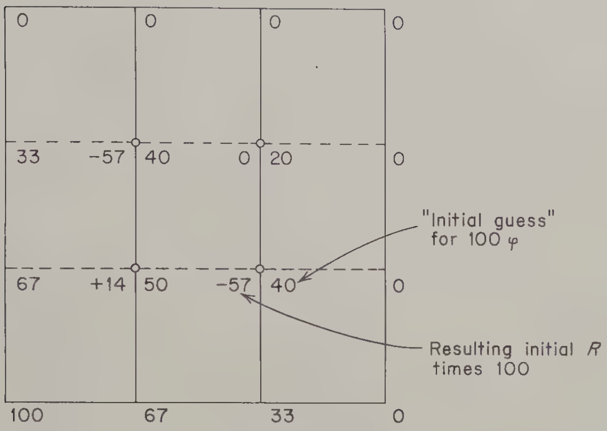
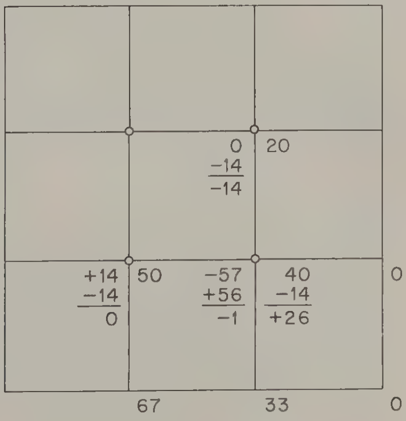


Figure 11.44

(−57) to −1. This is shown in Fig. 11.45, where corresponding changes in all the influenced residuals are indicated. It should be observed that there was no "carryover" to the external points.

Figure 11.46 is a complete "unit-relaxation solution," carried out in the manner just described. Convergence to the true solution is fairly rapid. It should be recognized, however, that for the case illustrated, the alternating + and − residuals tend to cancel each other. If the net sum of the residuals in an area is not equal to zero, these cancellations will not take place, and the solution will converge only after the residuals are "moved" by relaxation to the boundaries. This can be

Figure 11.45



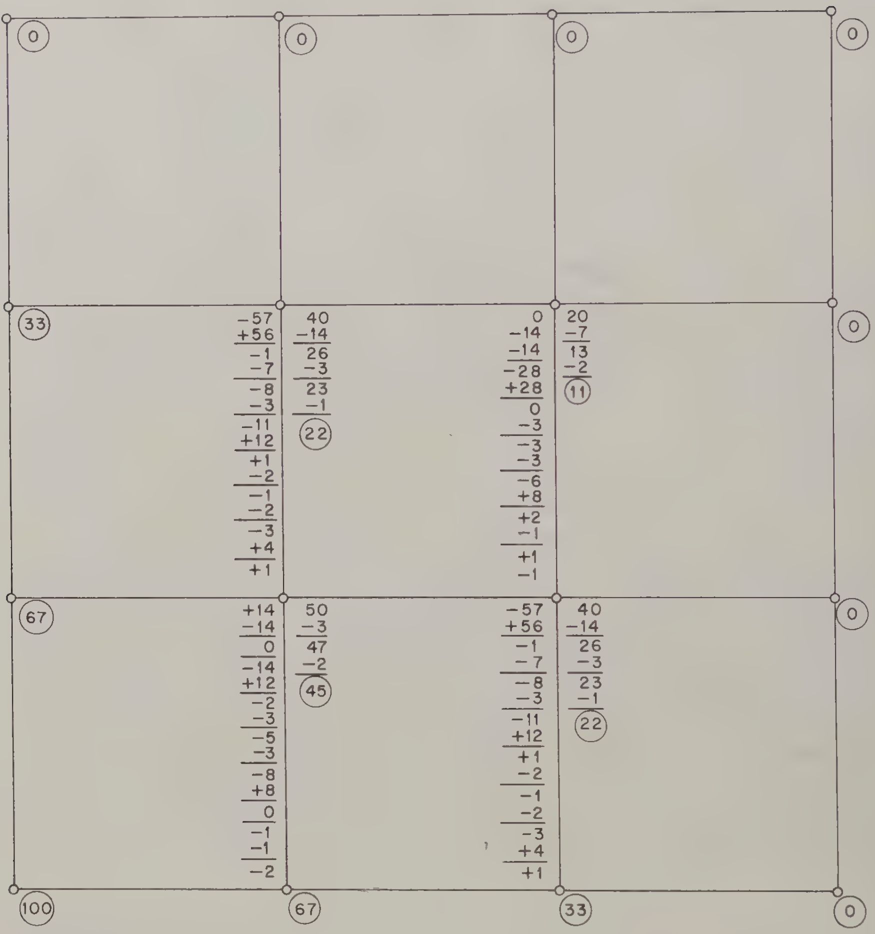


Figure 11.46

a lengthy process, since residuals change back and forth in sign but change little in magnitude. To circumvent this difficulty, group relaxation is often used.

11.6.2 Group Relaxation

As demonstrated in Example 11.1*b*, if φ at a given pivotal location is changed by a value of $+1$, the value of the residual at that location is altered by -4, whereas the values of the residuals at each of the adjacent pivotal locations are changed by $+1$. When φ values at several different, adjacent locations are changed, the relaxation pattern is altered. Con-

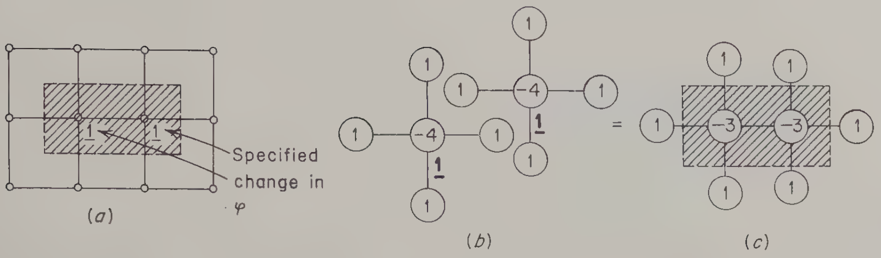


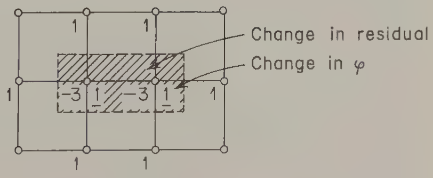
Figure 11.47 *(a)* Region in question; *(b)* combined Laplace operator; *(c)* resulting group operator.

sider, for example, the case illustrated in Fig. 11.47, where two adjacent points are altered simultaneously by a value of +1. The group-relaxation operator illustrated in Fig. 11.47*c* is frequently illustrated in the fashion of Fig. 11.48. Had three consecutive pivotal locations along a given grid line been subjected to changes in φ of 1, the group-relaxation operator would be that shown in Fig. 11.49. Figure 11.50 represents still another group-relaxation operator. Note especially the pattern at the reentrant point. It is also to be observed that the residual at the centermost point, where there are no grid connection lines to the "outside," is not changed during the group-relaxation process.

In comparing Figs. 11.48 to 11.50, certain common features are recognized. First, at each interior pivotal location—that is, interior to the region of group relaxation, where it is presumed that all φ values are altered by a value of +1—there will be a change in the residual at that point equal to −1 times the number of direct connections of that location to an exterior non-altered pivotal location. Secondly, at each "stationary" external location, there will be a change in the value of the residual equal to +1 times the number of direct connections to interior points.

Example 11.1c Example 11.1*b* could be group-relaxed as follows. The presumed starting values are those given in Fig. 11.51. If a relaxation that involves all the interior pivotal locations is assumed, the group

Figure 11.48



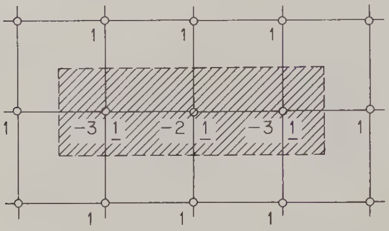


Figure 11.49

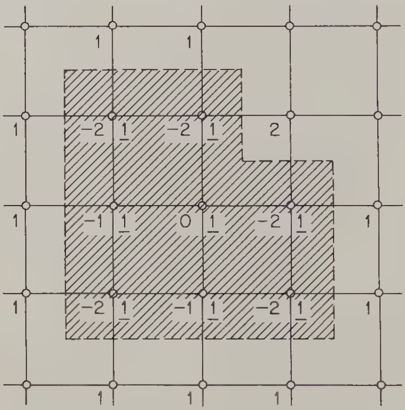
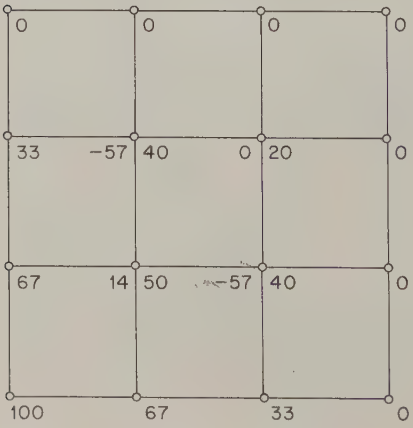


Figure 11.50

Figure 11.51



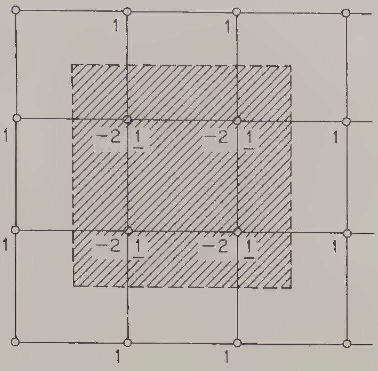


Figure 11.52

operator shown in Fig. 11.52 is obtained. It is to be remembered that any residual change moving into a boundary will simply be absorbed. There will be no change in the boundary value.

For the case in question (Fig. 11.51) the net residual within the selected interior region is

$$\Sigma R = -57 - 57 + 0 + 14 = -100 \tag{11.104}$$

It is desirable that this be removed—one part along each of the eight externally connecting paths. By selecting a change in φ at each interior pivotal location equal to -12, the corresponding change in R (at each of these interior locations) will be

$$R = (-2)(-12) = +24 \tag{11.105}$$

The total change within the region is $(4)(+24) = +96$. An initial group relaxation of the type just described results in the values shown in Fig. 11.53. From this stage, relaxation would normally proceed in the unit-relaxation fashion described in Example 11.1*b*.

Example 11.7 It is specified that within a rectangular region $d \times \frac{7}{3}d$ the laplacian

$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = 0 \tag{11.106}$$

must be satisfied. Moreover, the required boundary values are those listed in Fig. 11.54. It is to be noted that along part of the boundary φ is given, whereas, on the remainder, it is required that the normal derivative of φ equal zero.

The first step in the numerical solution of this problem is the selection

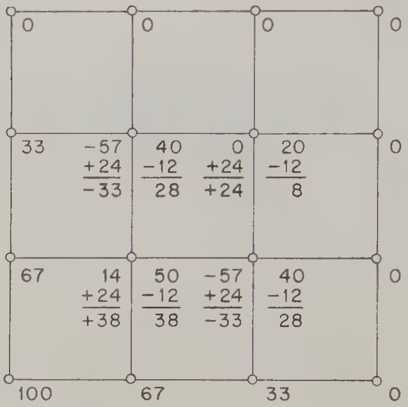


Figure 11.53

of a suitable grid. The chosen one is shown in Fig. 11.55. (Also included in this diagram are the imaginary pivotal locations required for the satisfaction of the specified boundary conditions.) The governing relaxation equations for the determination of the 19 unknown φ values are as follows:

$$
\begin{array}{r||l}
\text{At location } 2_a & \varphi_{2b} + 100 + \varphi_{2b} + 50 - 4\varphi_{2a} = R_{2a} \\
2_b & \varphi_{2a} + 100 + \varphi_{2c} + \varphi_{3b} - 4\varphi_{2b} = R_{2b} \\
2_c & \varphi_{2b} + 100 + \varphi_{2d} + \varphi_{3c} - 4\varphi_{2c} = R_{2c} \\
2_d & \varphi_{2c} + 100 + \varphi_{2e} + \varphi_{3d} - 4\varphi_{2d} = R_{2d} \\
2_e & \varphi_{2d} + 100 + \varphi_{2f} + \varphi_{3e} - 4\varphi_{2e} = R_{2e} \\
2_f & \varphi_{2e} + 100 + \varphi_{2g} + \varphi_{3f} - 4\varphi_{2f} = R_{2f} \\
2_g & \varphi_{2f} + 100 + 100 + \varphi_{3g} - 4\varphi_{2g} = R_{2g} \\
3_b & 50 + \varphi_{2b} + \varphi_{3c} + \varphi_{4b} - 4\varphi_{3b} = R_{3b} \\
3_c & \varphi_{3b} + \varphi_{2c} + \varphi_{3d} + \varphi_{4c} - 4\varphi_{3c} = R_{3c} \\
3_d & \varphi_{3c} + \varphi_{2d} + \varphi_{3e} + \varphi_{4d} - 4\varphi_{3d} = R_{3d} \\
3_e & \varphi_{3d} + \varphi_{2e} + \varphi_{3f} + \varphi_{4e} - 4\varphi_{3e} = R_{3e} \\
3_f & \varphi_{3e} + \varphi_{2f} + \varphi_{3g} + \varphi_{4f} - 4\varphi_{3f} = R_{3f} \\
3_g & \varphi_{3f} + \varphi_{2g} + 100 + \varphi_{4g} - 4\varphi_{3g} = R_{3g} \\
4_b & 50 + \varphi_{3b} + \varphi_{4c} + \varphi_{3b} - 4\varphi_{4b} = R_{4b} \\
4_c & \varphi_{4b} + \varphi_{3c} + \varphi_{4d} + \varphi_{3c} - 4\varphi_{4c} = R_{4c} \\
4_d & \varphi_{4c} + \varphi_{3d} + \varphi_{4e} + \varphi_{3d} - 4\varphi_{4d} = R_{4d} \\
4_e & \varphi_{4d} + \varphi_{3e} + \varphi_{4f} + \varphi_{3e} - 4\varphi_{4e} = R_{4e} \\
4_f & \varphi_{4e} + \varphi_{3f} + \varphi_{4g} + \varphi_{3f} - 4\varphi_{4f} = R_{4f} \\
4_g & \varphi_{4f} + \varphi_{3g} + 100 + \varphi_{3g} - 4\varphi_{4g} = R_{4g}
\end{array}
\qquad (11.107)
$$

(It is to be understood that, while these equations have been written in standard relaxation form, the set can be solved by any of the methods defined in Chap. 5.)

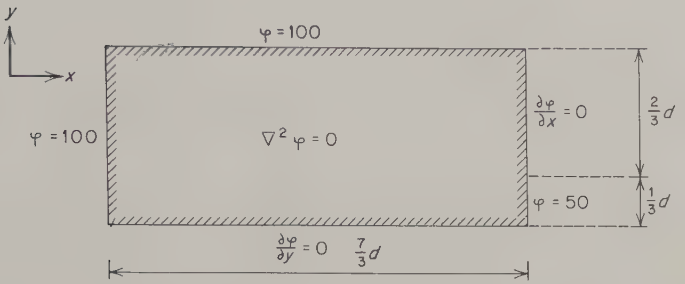


Figure 11.54

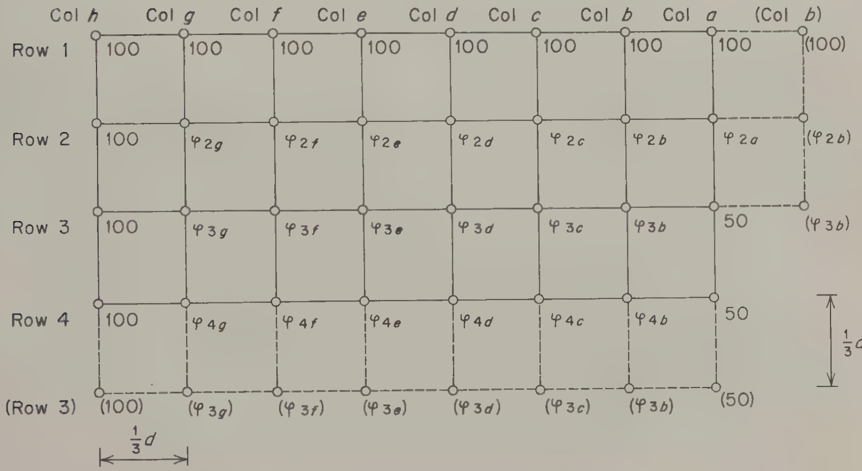
As a first approximation to the values of φ at the interior pivotal locations, the ones given in Fig. 11.56 will be presumed. The corresponding residuals are also listed. These estimates can be improved and convergence of the numerical process speeded up by going through a group relaxation for a part of the total region in question. The part to be examined is shown in Fig. 11.57. The group-relaxation operator is given in Fig. 11.58. From Fig. 11.57 it is observed that the net sum of the residuals within the region in question is

$$\Sigma R = 10 + 10 + 10 + 10 + 30 + 10 = 80 \tag{11.108}$$

This must be dissipated in equal amounts along the 12 externally connecting lines. Therefore

$$R_{\text{change along each external line}} = 80/12 \cong 7 \tag{11.109}$$

Figure 11.55



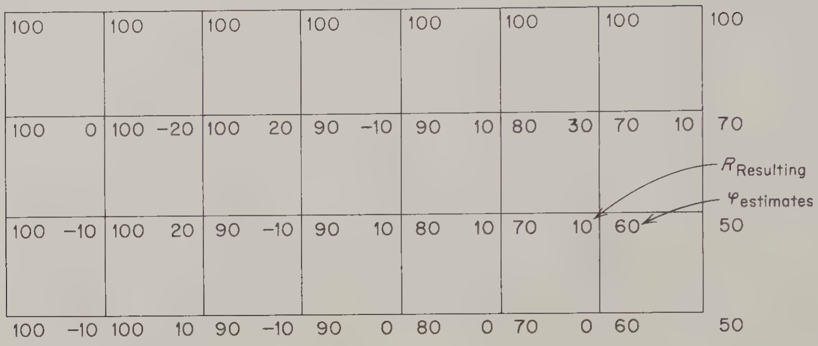


Figure 11.56

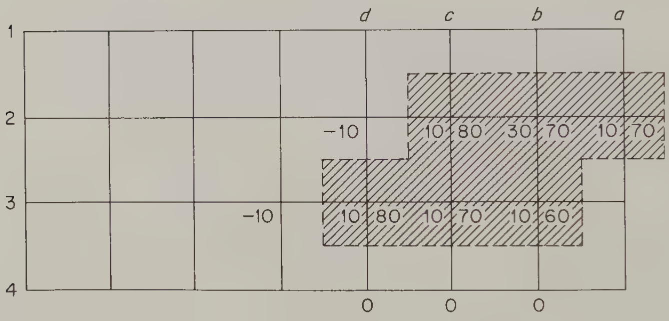
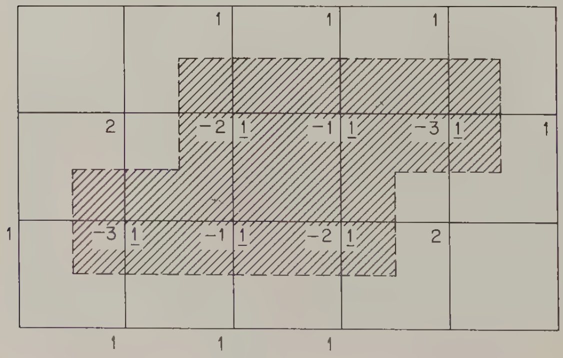
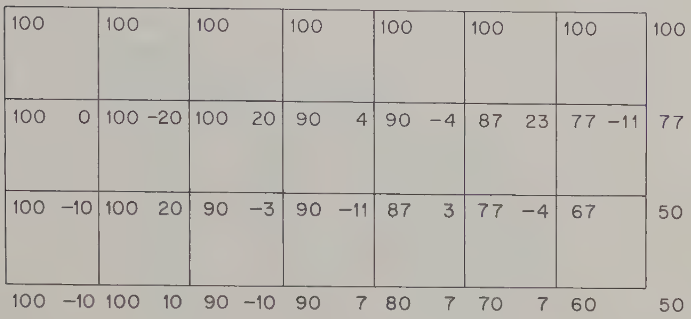


Figure 11.57

Figure 11.58 Group-relaxation operator.





100	100	100	100	100	100	100	100
100 0	100 −20	100 20	90 4	90 −4	87 23	77 −11	77
100 −10	100 20	90 −3	90 −11	87 3	77 −4	67	50
100 −10	100 10	90 −10	90 7	80 7	70 7	60	50

Figure 11.59

There results the modified form for Fig. 11.56 shown in Fig. 11.59. The unit-relaxation table for the total problem is given as Table 11.9. The final values of φ and R obtained by unit relaxation are given in Fig. 11.60. Contour lines of the electrical analog solutions to this problem are also indicated for purposes of comparison.

The relatively large errors that exist between the obtained numerical and the analog solutions are due almost entirely to the size of the selected grid in areas where φ must change rapidly. This effect can be corrected by completely reworking the problem using a finer mesh, or—as will be demonstrated—reworking the problem in the "troubled area" only, using the finer mesh and starting from the values obtained by using a coarse net.

11.6.3 Change in Grid Size

Example 11.7 illustrates the fact that even for one given problem, depending on the boundary conditions imposed, various degrees of accuracy may be desired. For a uniform spacing, the required accuracy in the most critical region dictates the entire pattern, and there results a large increase in the number of equations to be solved. A better procedure would be to use a varying grid size—one which allows a finer mesh in regions of rapid change, and coarser spacing where the variation is gradual.

To facilitate solution, if adjacent regions of unequal grid size are presumed, there must be available finite-difference operators which bridge the gap between the coarse and fine nets. If patterns having spatial dimensions related by factors of 2 are presumed, the possible types of interfaces are those illustrated in Fig. 11.61. Five points within the region are noted. These result in the five special spatial configurations shown in Fig. 11.62. It is to be recognized that equal-leg cases

Table 11.9

Unit Change in φ																			Resulting Change in Residuals																		
2a	*2b*	*2c*	*2d*	*2e*	*2f*	*2g*	*3b*	*3c*	*3d*	*3e*	*3f*	*3g*	*4b*	*4c*	*4d*	*4e*	*4f*	*4g*	*2a*	*2b*	*2c*	*2d*	*2e*	*2f*	*2g*	*3b*	*3c*	*3d*	*3e*	*3f*	*3g*	*4b*	*4c*	*4d*	*4e*	*4f*	*4g*
1																			4	1																	
	1																		1	4	1					1											
		1																		1	4	1					1										
			1																		1	4	1					1									
				1																		1	4	1					1								
					1																		1	4	1					1							
						1																		1	4	1					1						
							1													1					1	4	1					2					
								1													1					1	4	1					2				
									1													1					1	4	1					2			
										1													1					1	4	1					2		
											1													1					1	4	1					2	
												1													1					1	4	1					2
													1													1					1	4	1				
														1													1					1	4	1			
															1													1					1	4	1		
																1													1					1	4	1	
																	1													1					1	4	1
																		1													1					1	4

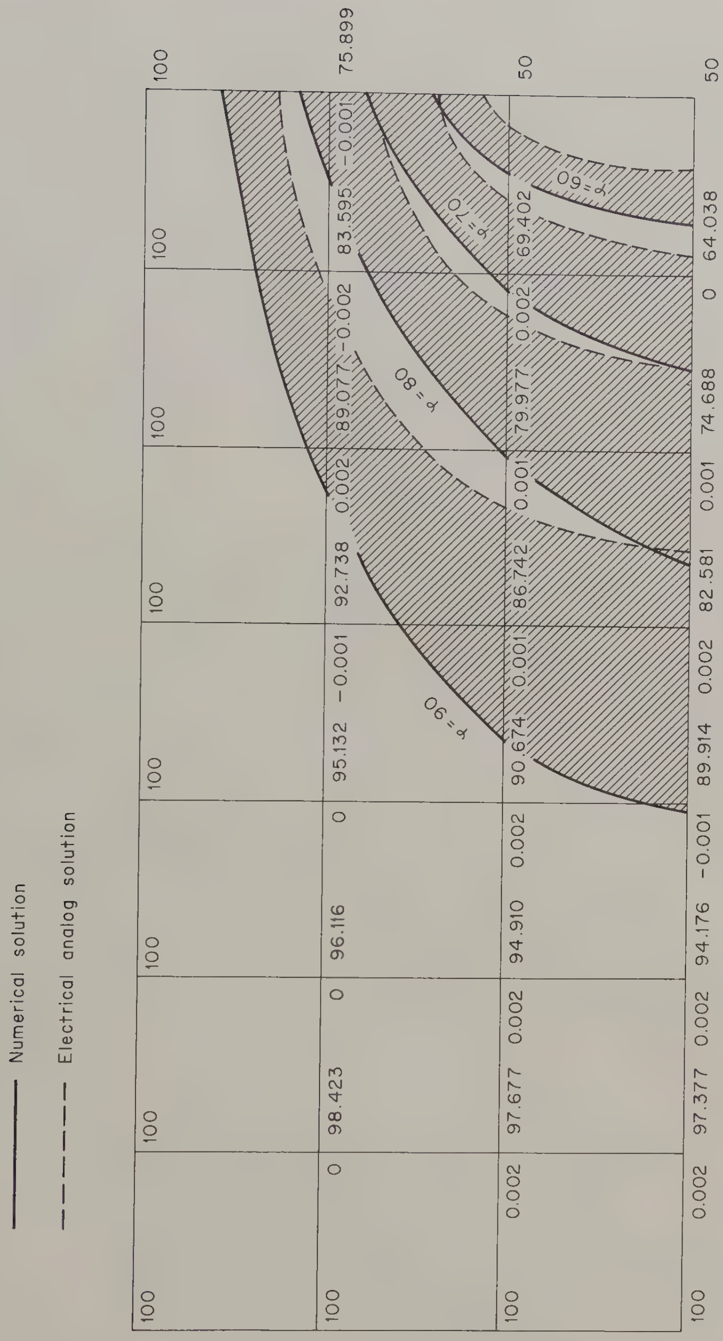

Numerical solution
Electrical analog solution
100 100 100 100 100 100 100 100
100 0 98.423 0 96.116 0 95.132 −0.001 92.738 0.002 89.077 −0.002 83.595 −0.001 75.899
100 0.002 97.677 0.002 94.910 0.002 90.674 0.001 86.742 0.001 79.977 0.002 69.402 50
100 0.002 97.377 0.002 94.176 −0.001 89.914 0.002 82.581 0.001 74.688 0 64.038 50
ϑ = 90
ϑ = 80
ϑ = 70
ϑ = 60

Figure 11.60

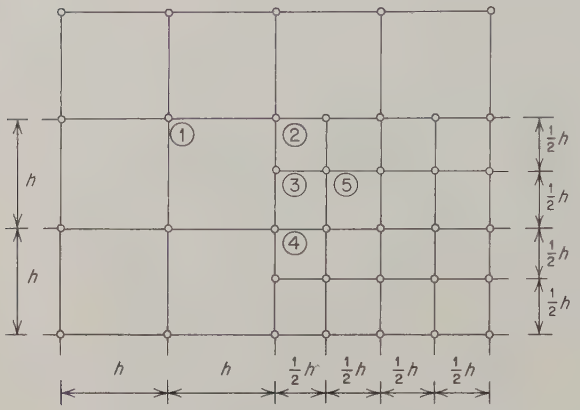
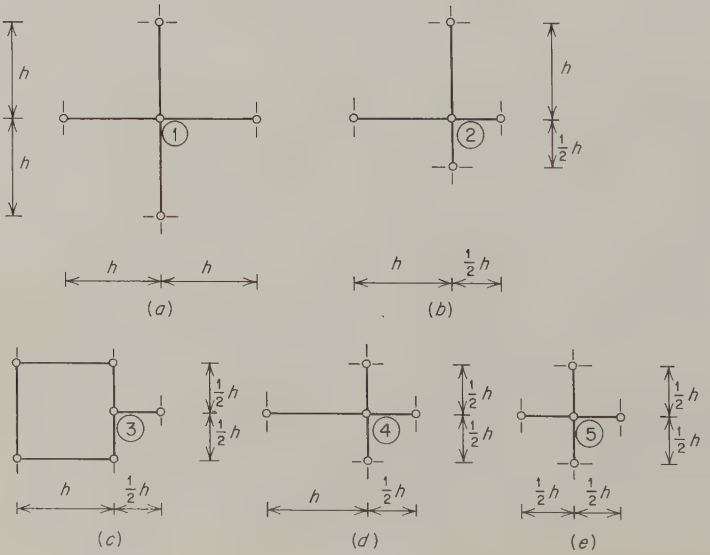


Figure 11.61

a and *e*, that is, configurations ① and ⑤, are of the same type, the grid size being the only difference. Points ② and ④ represent unequal-leg stars whose solutions to these are available in Sec. 11.4 (Fig. 11.17). Points of the configuration ③ type, i.e., case *c*, however, require special attention.

To establish the molecule associated with case *c*, the coordinate system

Figure 11.62



shown in Fig. 11.63 is assumed. It is further defined at the outset that variations in φ within the region in question ($1\frac{1}{2}h \times h$) can be represented to a sufficient degree of accuracy by the polynomial

$$\varphi = ax + by + cxy + dx^2 + ey^2 + \varphi_0 \tag{11.110}$$

The coefficients of this equation can be written in terms of the function values at pivotal locations 1 through 5 as follows:

$$\begin{bmatrix} -h & \frac{1}{2}h & -\frac{1}{2}h^2 & h^2 & \frac{1}{4}h^2 \\ & \frac{1}{2}h & & & \frac{1}{4}h^2 \\ \frac{1}{2}h & & & \frac{1}{4}h^2 & \\ & -\frac{1}{2}h & & & \frac{1}{4}h^2 \\ -h & -\frac{1}{2}h & \frac{1}{2}h^2 & h^2 & \frac{1}{4}h^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} + \begin{bmatrix} \varphi_0 \\ \varphi_0 \\ \varphi_0 \\ \varphi_0 \\ \varphi_0 \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \end{bmatrix} \tag{11.111}$$

Solution of this set of simultaneous equations gives, for example,

$$d = \frac{1}{3h^2}(-4\varphi_0 + \varphi_1 - \varphi_2 + 4\varphi_3 - \varphi_4 + \varphi_5) \tag{11.112}$$

$$e = \frac{2}{h^2}(\varphi_2 - 2\varphi_0 + \varphi_4) \tag{11.113}$$

If the computational molecule for the laplacian is desired, that is,

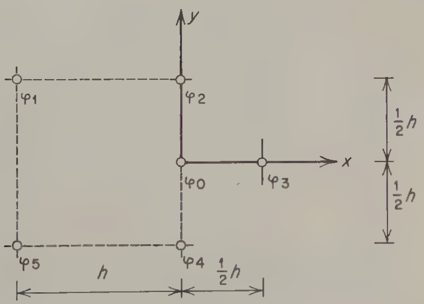
$$\nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} \tag{11.114}$$

the necessary derivatives of Eq. (11.110) are

$$\frac{\partial^2\varphi}{\partial x^2} = 2d \qquad \frac{\partial^2\varphi}{\partial y^2} = 2e \tag{11.115}$$

Substituting these into Eq. (11.114), and in turn Eqs. (11.112) and

Figure 11.63

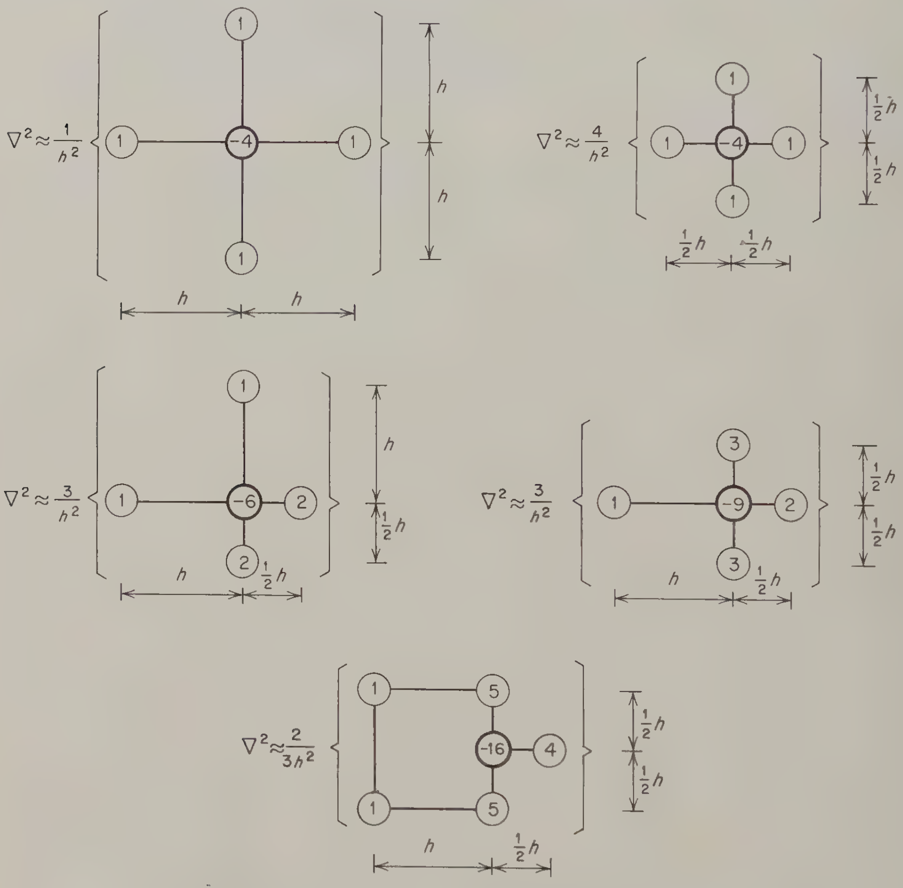


(11.113) into the resulting expression, yields

$$\nabla^2\varphi = \frac{1}{3h^2}(-32\varphi_0 + 2\varphi_1 + 10\varphi_2 + 8\varphi_3 + 10\varphi_4 + 2\varphi_5) \qquad (11.116)$$

Table 11.10 is a summary of the molecules required for solution of problems whose assumed unequal grid sizes have spatial dimensions related by a factor of 2.

Table 11.10 Laplacian Computational Molecules (Unequal Grids, Error of the Order of h^2)



With regard to the general question of what relative proportions of the total region in question should be subdivided, no clear rules exist. One could, for example, presume only a "unit" width of subdivided region in the area of anticipated greatest rate of change. For the example

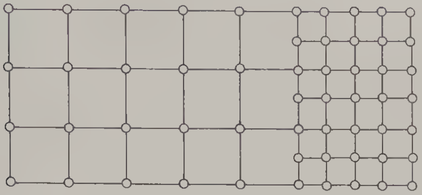


Figure 11.64

just considered, such an assumption would not result in any major computational difficulties, and the entire region could be considered in solving the problem. Then again, possibly two widths (or lengths) should have been presumed, as illustrated in Fig. 11.64. The resulting 39 unknowns could undoubtedly be handled. However, for other cases where larger regions are considered, it may be desirable to presume that the function values obtained from the coarser-grid analysis are constants at distances of three, four, or five grid spacings from the boundaries of the subdivided region.

Example 11.8 To illustrate the procedure for mixed-grid problems, consider the region defined in Fig. 11.65. The rows have been numbered ① through ⑥ and the columns ⓐ through ⓔ. Interior point types corresponding to those illustrated in Fig. 11.62 and Table 11.10 are also shown, as are the presumed, constant boundary conditions.

Application of the proper operator at each of the interior grid points gives rise to a set of 18 equations in 18 unknown values of φ. The coefficients of these equations are given in Table 11.11. It should be recognized that the transpose of this coefficient matrix is the unit-relaxation table. Contour lines of the resulting solution are indicated in Fig.

Figure 11.65

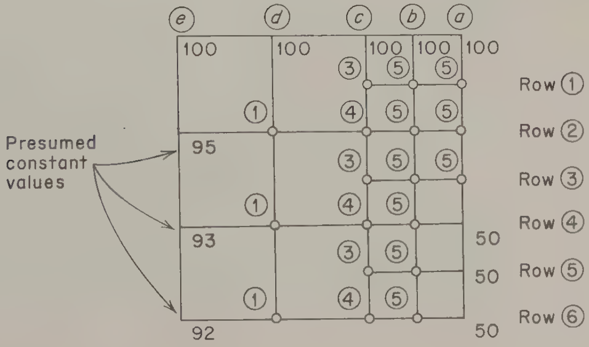


Table 11.11

1*a*	1*b*	1*c*	2*a*	2*b*	2*c*	2*d*	3*a*	3*b*	3*c*	4*b*	4*c*	4*d*	5*b*	5*c*	6*b*	6*c*	6*d*	= Const.
−4	2		1															−100
1	−4	1		1														−100
	4	−16			5	1												−600
1			−4	2			1											0
	1		1	−4	1			1										0
		3.08		2	−9.23	1			3.08									0
					1	−4						1						−195
			1				−4	2										−50
				1			1	−4	1	1								0
					5	1		4	−16		5	1						0
								1		−4	1		1					−50
									3.08	2	−9.23	1		3.08				0
						1					1	−4					1	−93
										1			−4	1	1			−50
											5	1	4	−16		5	1	0
													2		−4	1		−50
														6.16	2	−9.23	1	0
												2				1	−4	−92

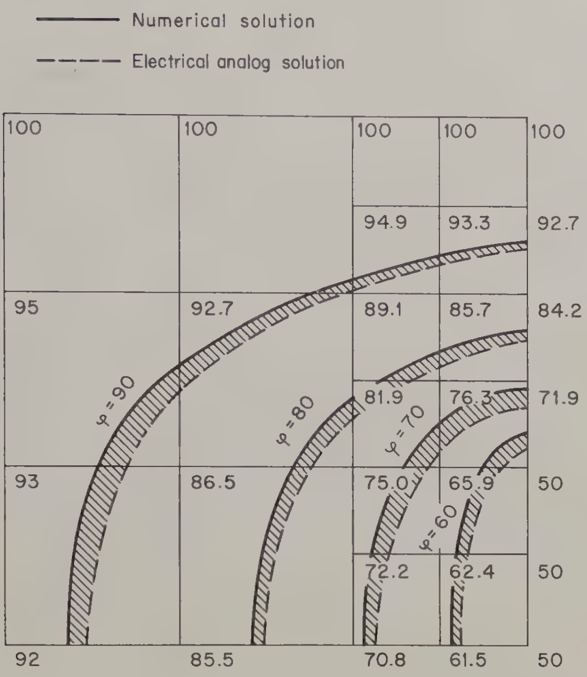


Figure 11.66

11.66 by the solid lines, and by the values listed at each pivotal location. "More exact" solutions, obtained from an electrical analog of the problem, are shown dashed.

11.7 Parabolic Partial Differential Equations: Transient Problems

If convection and radiation are neglected, the temperature distribution along a uniform-cross-section wire of length L satisfies the parabolic partial differential equation

$$\frac{\partial u}{\partial t} = K\frac{\partial^2 u}{\partial x^2} \tag{11.117}$$

where $u(x,t)$ = temperature of bar
K = thermal diffusivity

Initial and boundary conditions for such problems are frequently of the type

$$\begin{aligned} u(0,t) &= f_1(t) \\ u(L,t) &= f_2(t) \\ u(x,0) &= g(x) \end{aligned} \tag{11.118}$$

The corresponding two-dimensional space-time heat-conduction equation is

$$\frac{\partial u}{\partial t} = K\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{11.119}$$

In Secs. 11.5 and 11.6 it was demonstrated that for elliptic-type equations the value of the solution at any one interior point depends on *all* of the boundary values. Such is not the case for parabolic-type problems. Here, by assuming for illustration the one-dimensional space-time problems, the value of $u(x_i,t_i)$ depends only on the values of $f_1(t)$ and $f_2(t)$ in the range $0 \le t \le t_i$. This means that in the solution of these types of problems, a step-by-step or "marching" method can be employed. On the surface, then, it would appear that parabolic-type partial differential equations should be easier to deal with than elliptic types. Unfortunately, this is not necessarily the case. Marching methods, by their very nature, introduce a difficulty not usually present in elliptic problems, namely, numerical "instability." However, as will be demonstrated, these special problems can be overcome by a proper choice of the nodal increments.

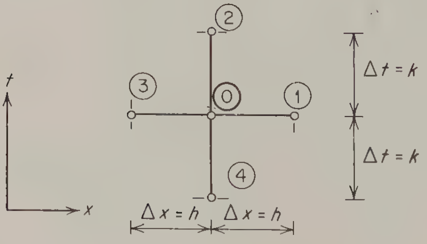
11.7.1 Explicit Formulation

The essential features of the finite-difference procedures developed for elliptic equations can be extended to the solution of parabolic equations. For example, for the two-dimensional case (one space dimension and one time dimension), the grid system of Fig. 11.67 can be used. If it is assumed that the partial differential equation in question is

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{11.120}$$

difference approximations for the first and second derivatives are required. If the second-order central differences in both of the coordinate directions are desired, the necessary equations (from Table 9.27) are the

Figure 11.67



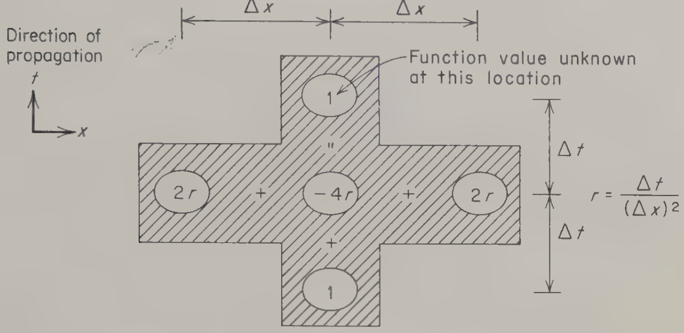


Figure 11.68

following:

$$\frac{\partial u}{\partial t} \cong \frac{1}{2k}\,(-u_4 + u_2) \tag{11.121}$$

$$\frac{\partial^2 u}{\partial x^2} \cong \frac{1}{h^2}\,(u_3 - 2u_0 + u_1) \tag{11.122}$$

Substitution of these into the given differential equation yields

$$\frac{1}{2k}\,(-u_4 + u_2) = \frac{1}{h^2}\,(u_3 - 2u_0 + u_1) \tag{11.123}$$

By solving for u_2, the *explicit* equation for continuing in the t direction a numerical solution would be

$$u_2 = 2r\,(u_3 - 2u_0 + u_1) + V_4 \tag{11.124}$$

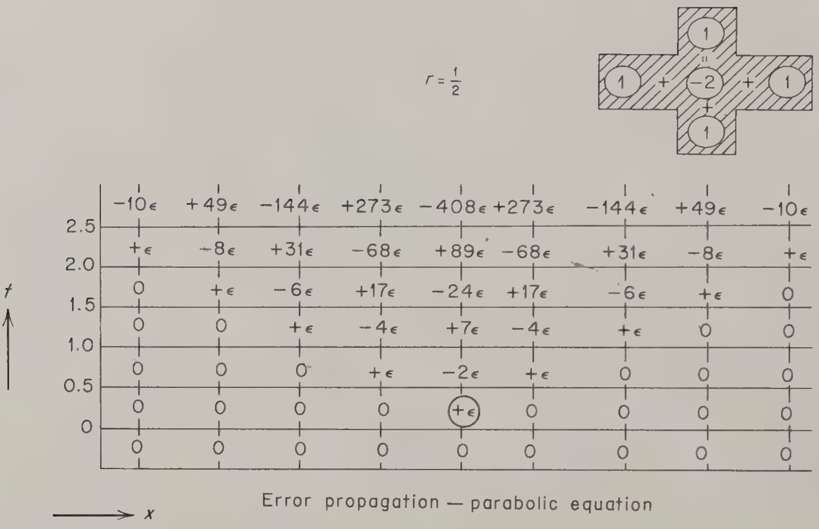
where

$$r = \frac{k}{h^2} = \frac{\Delta t}{(\Delta x)^2} \tag{11.125}$$

The corresponding computational molecule is shown in Fig. 11.68.

To examine the suitability of this molecule, consider the propagation of an error ϵ, presumed to be introduced in the computational process at a single pivotal location. This error could be due to round-off or other reasons. Two cases, $r = \frac{1}{2}$ and $r = 1.0$, will be examined. For both, to allow a comparison of the resulting values, Δx is presumed equal to 1.0. Clearly, from Figs. 11.69 and 11.70, the computational molecule defined by Fig. 11.68 is *numerically unstable,* for both $r = 1.0$ and $r = \frac{1}{2}$. Moreover, it is interesting to note that the magnitude of the error increases for a decrease in the grid spacing in the t direction.

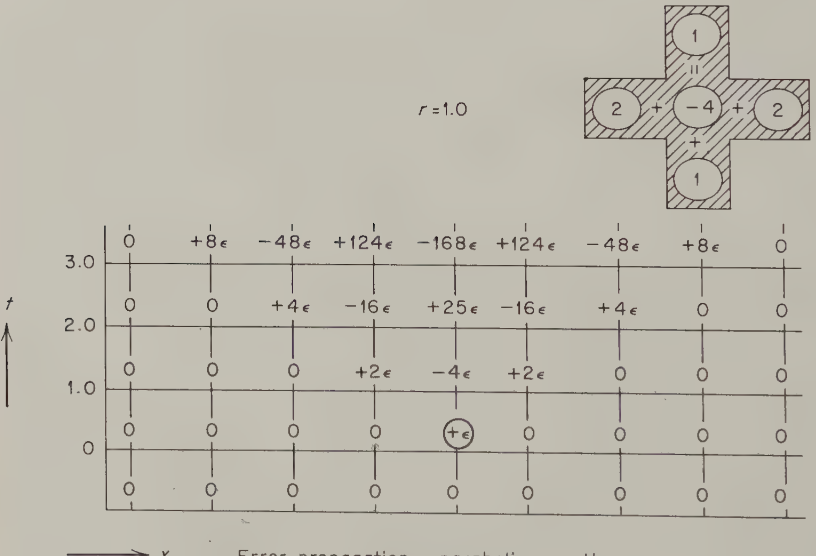
Instead of presuming the central-finite-difference expression [Eq. (11.121)] for the first derivative of u with respect to t, it will now be assumed that the most simple forward-difference expression is to be used.


$r = \frac{1}{2}$
1
1 + −2 + 1
1
t
x
2.5
2.0
1.5
1.0
0.5
0
−10ε +49ε −144ε +273ε −408ε +273ε −144ε +49ε −10ε
+ε −8ε +31ε −68ε +89ε −68ε +31ε −8ε +ε
0 +ε −6ε +17ε −24ε +17ε −6ε +ε 0
0 0 +ε −4ε +7ε −4ε +ε 0 0
0 0 0 +ε −2ε +ε 0 0 0
0 0 0 0 +ε 0 0 0 0
0 0 0 0 0 0 0 0 0

Error propagation — parabolic equation

Figure 11.69

Figure 11.70



Error propagation — parabolic equation

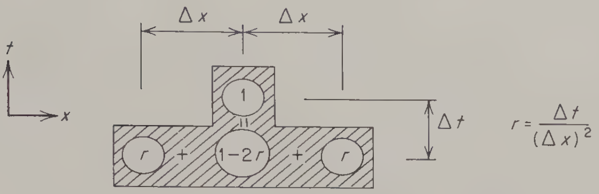


Figure 11.71

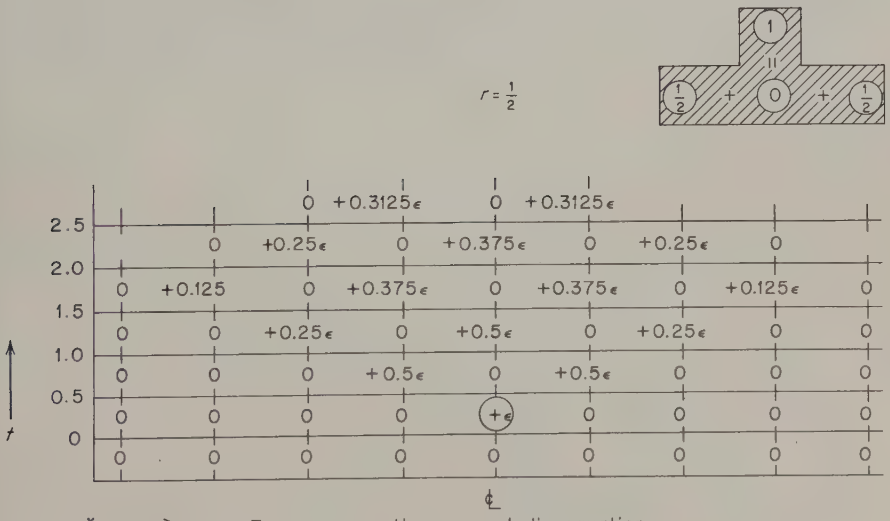
That is,

$$\frac{\partial u}{\partial t} \cong \frac{1}{k}(-u_0 + u_2) \tag{11.126}$$

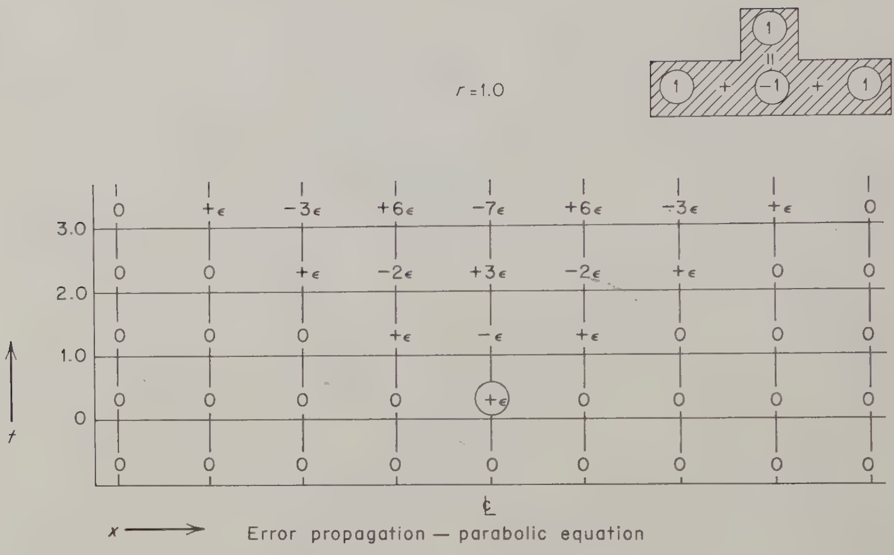
Again using Eq. (11.123) for the definition of the second derivative of u with respect to x gives the computational molecule shown in Fig. 11.71. The error propagation for this molecule is illustrated in Figs. 11.72 and 11.73 for $r = \frac{1}{2}$ and 1.0, respectively. For $r = \frac{1}{2}$ the error decreases with t. For $r = 1.0$ an increase is observed.

Possibly a clearer indication of the error propagation associated with the molecule shown in Fig. 11.71 can be seen from a plot of the absolute values of the errors at the x location where the error was first introduced. (It is to be recognized that while the values shown on such a plot repre-

Figure 11.72


r = 1/2
1
=
1/2
+
0
+
1/2
0 +0.3125ε 0 +0.3125ε
2.5
0 +0.25ε 0 +0.375ε 0 +0.25ε 0
2.0
0 +0.125 0 +0.375ε 0 +0.375ε 0 +0.125ε 0
1.5
0 0 +0.25ε 0 +0.5ε 0 +0.25ε 0 0
1.0
0 0 0 +0.5ε 0 +0.5ε 0 0 0
0.5
0 0 0 0 +ε 0 0 0 0
0
0 0 0 0 0 0 0 0 0
t
x
℄

Error propagation — parabolic equation



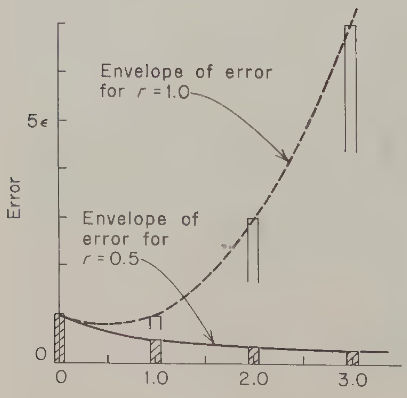
Error propagation — parabolic equation

Figure 11.73

sent discontinuous points, an "envelope of errors" can be drawn for each of the cases considered.)

Figures 11.72 and 11.74 do not prove, but only illustrate, that the simple forward finite-difference expression with $r = \frac{1}{2}$ yields a stable solution. It can be shown analytically, however, that solutions to parabolic partial differential equations obtained by using the computational molecule of Fig. 11.71 will converge to the correct solution providing that r is fixed and $r \leq \frac{1}{2}$. Moreover, it can be shown that such solutions

Figure 11.74



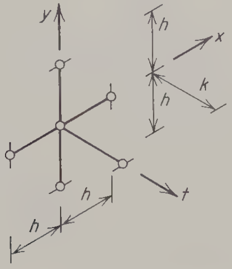


Figure 11.75

are stable.† A further advantage to selecting $r = \frac{1}{2}$ can be seen in the computational molecule of Fig. 11.72. The multiplier for the i, j position is zero. Moreover, the value $\varphi(i, j + \Delta t)$ is the average of the two values $\varphi(i - \Delta h, j)$ and $\varphi(i + \Delta h, j)$.

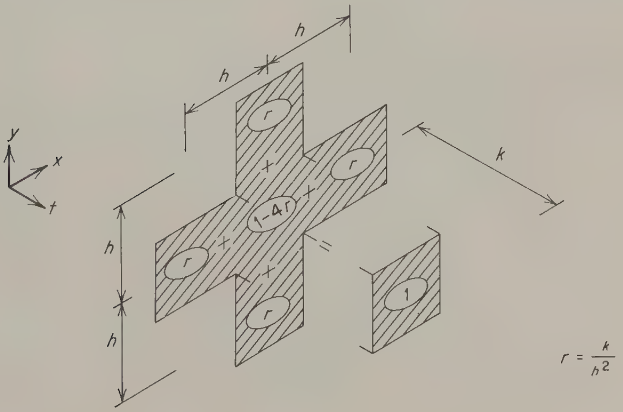
Computational molecules for use in the solution of three-dimensional parabolic problems (two spatial dimensions and one time dimension) can be obtained in a manner similar to that developed above. For the grid system shown in Fig. 11.75, the computational molecule is given in Fig. 11.76. For stability and convergence it is required that

$$r = \frac{k}{h^2} \leq \frac{1}{4} \tag{11.127}$$

---

† A solution is said to converge if the difference between the approximation and the exact solution approaches zero as the size of the increment used approaches zero. A solution is stable if an error from any source diminishes as time increases.

Figure 11.76



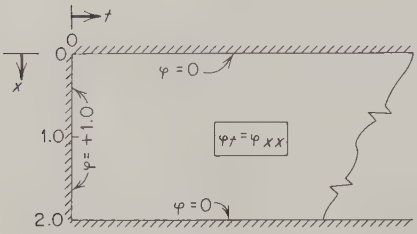


Figure 11.77

Example 11.9 To illustrate the application of explicit-recurrence techniques, consider the numerical solution of the parabolic partial differential equation problem defined by Eqs. (11.128) and (11.129).

$$\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} \tag{11.128}$$

$$\varphi(0,t) = 0 \qquad \varphi(2,t) = 0 \qquad \varphi(x,0) = 1.0 \tag{11.129}$$

The solution domain is shown in Fig. 11.77. It is to be noted that since the initial condition $\varphi(x,0)$ does not vary with x, and since both "lateral boundary conditions" are equal, symmetry of solution about $x = 1.0$ is to be expected. The problem then can be reduced to a consideration of the case illustrated in Fig. 11.78.

For the purpose of comparison, an analytical solution to the problem will be obtained first. If a product solution of the type defined by Eq. (11.130) is assumed,

$$\varphi = F(x)G(t) \tag{11.130}$$

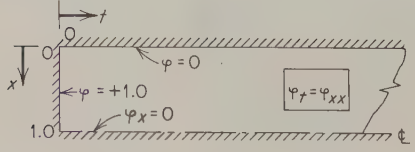
then the given differential equation reduces to the following two ordinary differential equations:

$$F'' = -\lambda^2 F \qquad \text{and} \qquad G' = -\lambda^2 G \tag{11.131}$$

The general solution will be, therefore,

$$\varphi = (C_1 \cos \lambda x + C_2 \sin \lambda x) C_3 e^{-\lambda^2 t} \tag{11.132}$$

Figure 11.78



To satisfy the boundary condition $\varphi(0,t) = 0$

$$C_1C_3 = 0 \tag{11.133}$$

To ensure that $\varphi'(1.0,t) = 0$, that is, to satisfy the second boundary condition,

$$\cos \lambda x = 0 \tag{11.134}$$

or

$$\lambda = \frac{n\pi}{2} \qquad (n = 1, 3, 5, \ldots) \tag{11.135}$$

For the initial condition $\varphi(x,0) = +1.0$ (denoting $C_2C_3 = C_4$),

$$\sum_{n=1,3,5,\ldots}^{\infty} C_4 \sin \frac{n\pi x}{2} = 1.0 \tag{11.136}$$

or

$$C_4 = 2 \int_0^1 \sin \frac{n\pi}{2} x \, dx$$

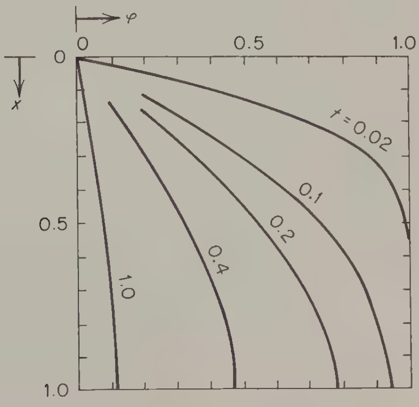
The analytical solution to the problem in question is, therefore,

$$\varphi = \sum_{n=1,3,5,\ldots}^{\infty} \frac{4}{n\pi} \left(\sin \frac{n\pi}{2} x \right) e^{-(n^2\pi^2/4)t} \tag{11.137}$$

The solution is shown graphically in Fig. 11.79.

It will first be assumed that a four-equal-segment grid spacing in the x direction will result in a sufficiently accurate numerical solution. That is, Δx will be selected equal to $\frac{1}{4}$. Quite arbitrarily, the ratio

Figure 11.79



$r = \Delta t/\Delta x^2$ will be taken as $\frac{1}{6}$. The resulting value of Δt is then

$$\Delta t = \left(\frac{1}{4}\right)^2 r = \left(\frac{1}{16}\right)\left(\frac{1}{6}\right) = 0.0104 \tag{11.138}$$

The grid system under examination is shown in Fig. 11.80. Also indicated is the corresponding explicit-recurrence computational molecule. It is to be noted that there is an apparent inconsistency in the listed initial (or boundary) value of φ at $x = 0$, $t = 0$. Along the x direction the required value is $\varphi = 1.0$. The boundary condition for all values of t at $x = 0$ requires that $\varphi = 0$. The average value has been listed.

An extra row of grid points outside the domain of interest has been introduced to facilitate the handling of the boundary condition

$$\varphi_x(1.0,t) = 0 \tag{11.139}$$

From Fig. 10.13, it should be evident that such a boundary condition requires that

$$\varphi(0.75,t) = \varphi(1.25,t) \tag{11.140}$$

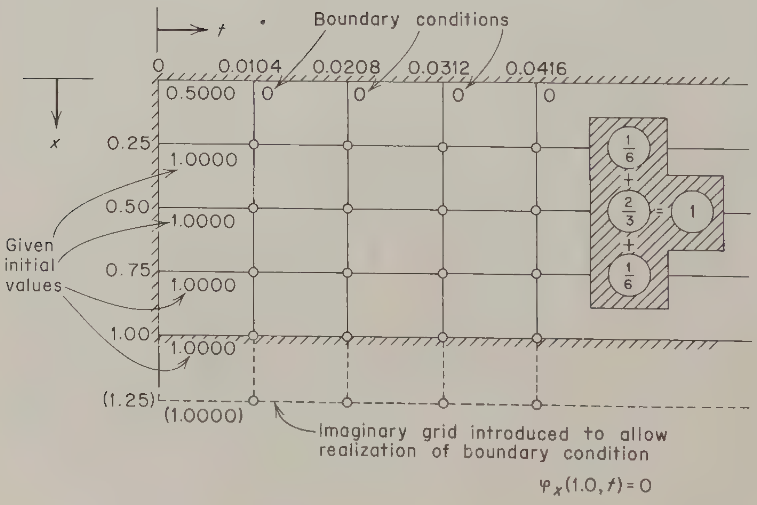
Figure 11.81 shows the computational molecule placed correctly for the evaluation of $\varphi(0.25,0.0104)$. The computation is that illustrated in Eq. (11.141).

$$[1]\,[\varphi(0.25,0.0104)] = \left(\frac{1}{6}\right)[0.5000] + \left(\frac{2}{3}\right)[1.0000] + \left(\frac{1}{6}\right)[1.0000] \tag{11.141}$$

or

$$\varphi(0.25,0.0104) = 0.9168 \tag{11.142}$$

Figure 11.80



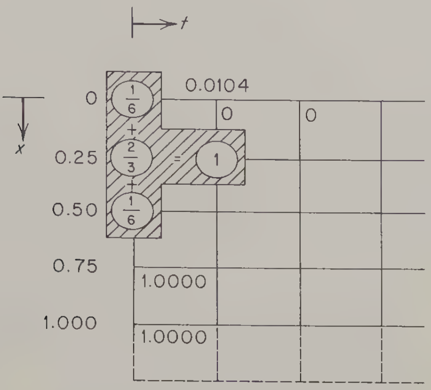


Figure 11.81

Values of φ for the range $0 \leq t \leq 0.1$ are summarized in Fig. 11.82. It is to be noted that to facilitate presentation, the coordinate directions in this figure have been interchanged. That is, the variable t has been listed vertically, and x horizontally. To give an indication of the accuracy of solution obtained, Table 11.12 compares the numerical values obtained from $t = 0.1040$ with corresponding values from the analytical solution [Eq. (11.137)].

Figure 11.82

t \ x	0	0.25	0.50	0.75	1.00	1.25
0	0.5000	1.0000	1.0000	1.0000	1.0000	1.0000
0.0104	0	0.9168	1.0000	1.0000	1.0000	1.0000
0.0208	0	0.7779	0.9862	1.0000	1.0000	1.0000
0.0312	0	0.6830	0.9539	0.9978	1.0000	0.9978
0.0416	0	0.6144	0.9162	0.9909	0.9994	0.9909
0.0520	0	0.5624	0.8784	0.9800	0.9966	0.9800
0.0624	0	0.5214	0.8427	0.9659	0.9911	0.9659
0.0728	0	0.4881	0.8098	0.9497	0.9827	0.9497
0.0832	0	0.4604	0.7796	0.9320	0.9717	0.9320
0.0936	0	0.4369	0.7519	0.9133	0.9585	0.9133
0.1040	0	0.4166	0.7268	0.8940	0.9434	0.8940

Table 11.12

x \ φ	0.25	0.50	0.75	1.00
φ Analytical	0.4157	0.7261	0.8938	0.9428
φ Numerical	0.4166	0.7268	0.8940	0.9434

Example 11.10 For the same problem, by selecting r equal to $\frac{1}{2}$ and maintaining Δx at 0.25, the increment in Δt is

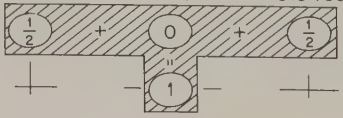
$$\Delta t = (\tfrac{1}{4})^2(\tfrac{1}{2}) = \frac{1}{32} = 0.03125 \tag{11.143}$$

Solution for this set of conditions is summarized in Fig. 11.83. Also shown is the corresponding computational molecule. A plot of these values is shown in Fig. 11.84.

Example 11.11 Little additional difficulty is introduced when the differential equation also contains the first derivative of the function

Figure 11.83

t \ x	0	0.25	0.50	0.75	1.00	1.25
0	0.5000	1.0000	1.0000	1.0000	1.0000	1.0000
0.0313	0	0.7500	1.0000	1.0000	1.0000	1.0000
0.0625	0	0.5000	0.8750	1.0000	1.0000	1.0000
0.0938	0	0.4375	0.7500	0.9375	1.0000	0.9375
0.1250	0	0.3750	0.6875	0.8750	0.9375	0.8750
0.1563	0	0.3438	0.6250	0.8125	0.8750	0.8125
0.1875	0	0.3125	0.5782	0.7500	0.8125	0.7500
0.2188	0	0.2891	0.5313	0.6954	0.7500	0.6954
0.2500	0	0.2657	0.4923	0.6406	0.6954	0.6406
0.2813	0	0.2462	0.4532	0.5939	0.6406	0.5939
0.3125	0	0.2266	0.4201	0.5469	0.5939	0.5469



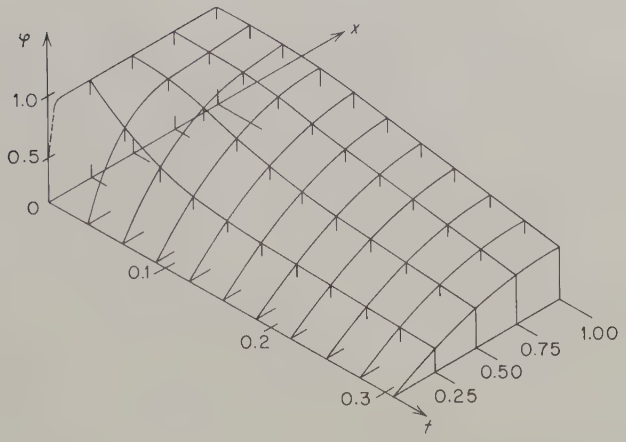


Figure 11.84

with respect to x. Consider, for example, the equation

$$\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} + 4\,\frac{\partial \varphi}{\partial x} \tag{11.144}$$

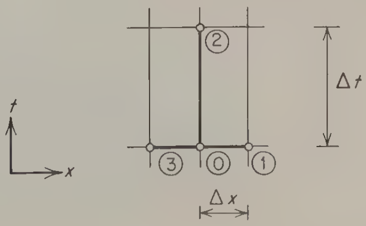
It will be presumed that the same boundary conditions hold as were prescribed for Example 11.9. That is,

$$\varphi(0,t) = 0 \qquad \varphi_x(1,t) = 0 \qquad \varphi(x,0) = 1.0 \tag{11.145}$$

By presuming the numbered grid system of Fig. 11.85, Eq. (11.144) can be written in different form, as given in Eq. (11.146).

$$\frac{1}{\Delta t}(-\varphi_0 + \varphi_2) = \frac{1}{(\Delta x)^2}(\varphi_3 - 2\varphi_0 + \varphi_1) + \frac{4}{2(\Delta x)}(-\varphi_3 + \varphi_1) \tag{11.146}$$

Figure 11.85



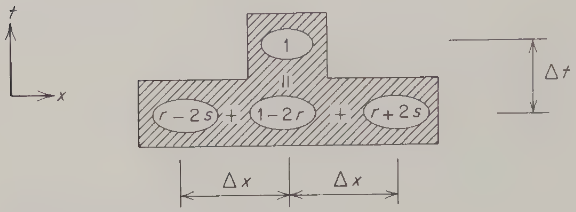


Figure 11.86

By defining

$$r = \frac{\Delta t}{(\Delta x)^2} \tag{11.147}$$

and

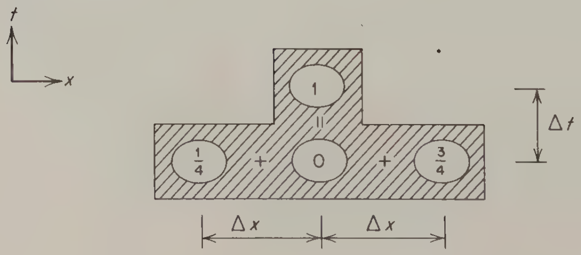
$$s = \frac{\Delta t}{\Delta x} \tag{11.148}$$

the computational molecule is that shown in Fig. 11.86. If r is selected equal to $\frac{1}{2}$ and $\Delta x = \frac{1}{4}$, then s will be equal to $\frac{1}{8}$, and the computational molecule will be that of Fig. 11.87. Error propagation is examined in Fig. 11.88. Solution would proceed in the manner of the preceding cases.

Example 11.12 To illustrate an "explicit" numerical solution of a parabolic equation in two spatial and one time coordinates, it is presumed that a spatial region $0 \leq x \leq 1.0$, $0 \leq y \leq 1.0$ is subjected to the initial condition defined in Fig. 11.89. The specific equation to be solved is

$$\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} \tag{11.149}$$

Figure 11.87



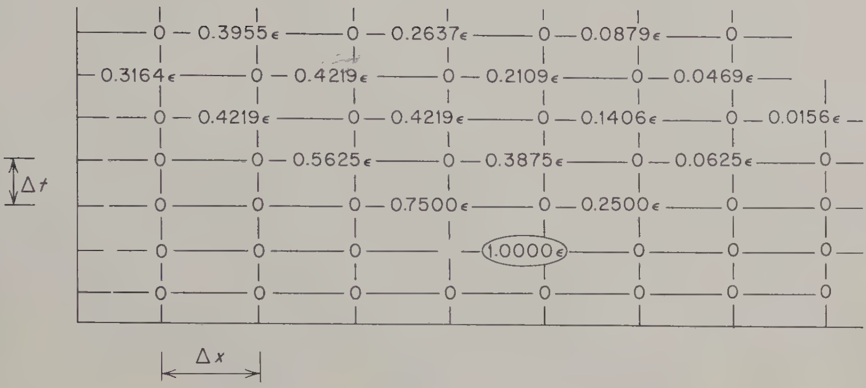

0 — 0.3955ε — 0 — 0.2637ε — 0 — 0.0879ε — 0
0.3164ε — 0 — 0.4219ε — 0 — 0.2109ε — 0 — 0.0469ε
0 — 0.4219ε — 0 — 0.4219ε — 0 — 0.1406ε — 0 — 0.0156ε
0 — 0 — 0.5625ε — 0 — 0.3875ε — 0 — 0.0625ε — 0
0 — 0 — 0 — 0.7500ε — 0 — 0.2500ε — 0 — 0
0 — 0 — 0 — 1.0000ε — 0 — 0 — 0
0 — 0 — 0 — 0 — 0 — 0 — 0 — 0
Δt
Δx

Figure 11.88

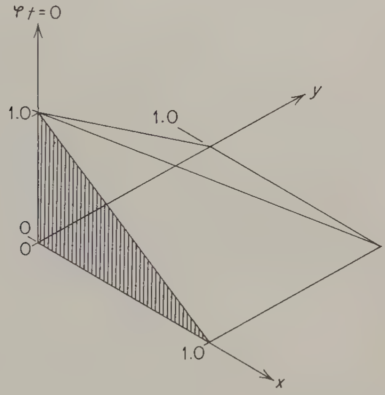
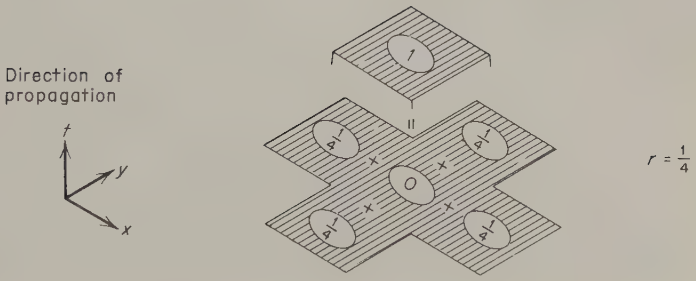


Figure 11.89

Figure 11.90



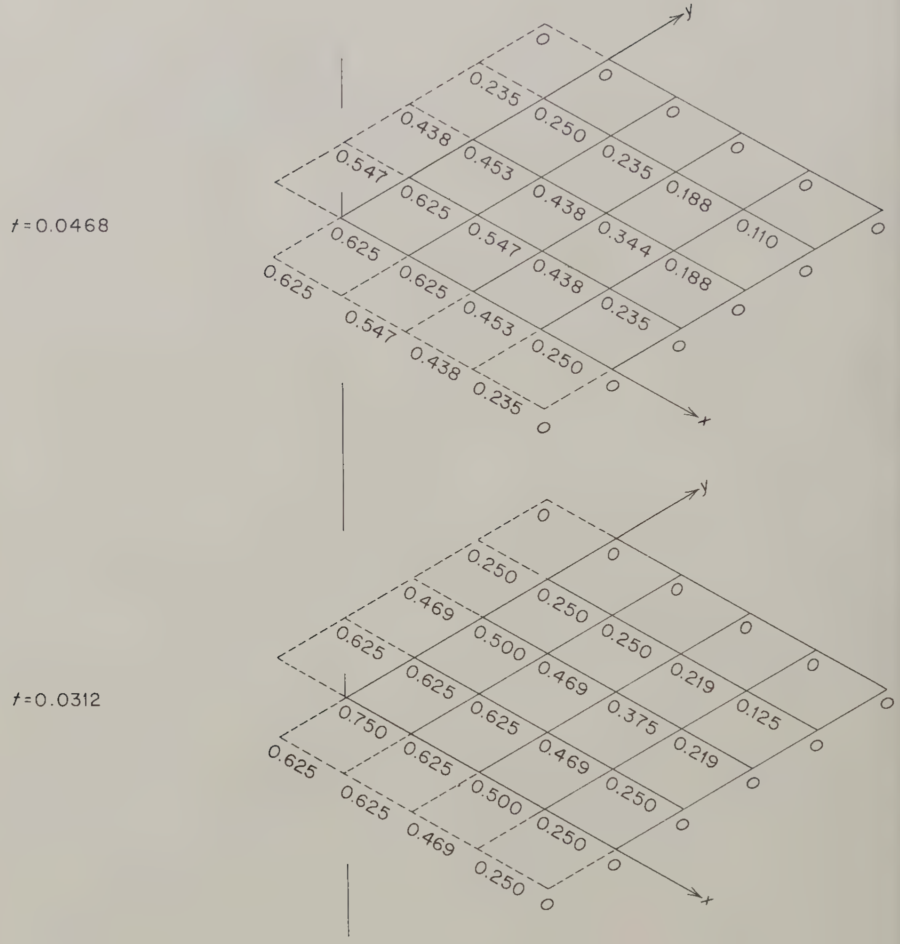
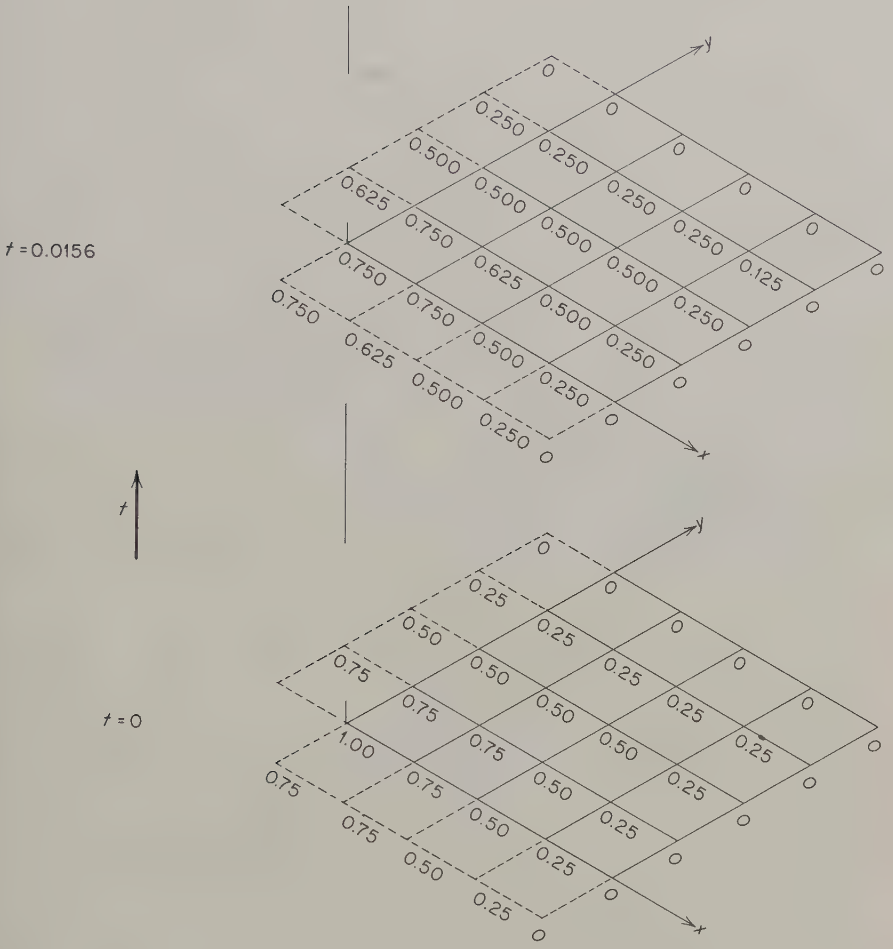


Figure 11.91

The boundary conditions for $t \neq 0$ are

$$\begin{aligned} \varphi_x(0,y,t) &= 0 \\ \varphi_y(x,0,t) &= 0 \\ \varphi(1.0,y,t) &= 0 \\ \varphi(x,1.0,t) &= 0 \end{aligned} \tag{11.150}$$

By choosing $\Delta x = \Delta y$ and $r = \frac{1}{4}$ (see Fig. 11.76), the computational molecule shown in Fig. 11.90 is obtained. Solutions for the first three increments in time are shown in Fig. 11.91.



11.7.2 Implicit Formulation

The explicit method of solution described in Sec. 11.7.1 has two major drawbacks. First, to ensure stability, only very small increments in the *time variable* are permissible. This frequently leads to excessive computation. Secondly, as is evident from a consideration of any of the previously developed explicit computational molecules, $\varphi_{x,t+\Delta t}$ is only dependent on $\varphi_{x-\Delta x,t}$, $\varphi_{x,t}$, and $\varphi_{x+\Delta x,t}$. The boundary conditions at $t = 1$ are not involved. This is illustrated in Fig. 11.92. A more correct representation would require that the definition of φ at any particular value of time, regardless of the x value of the pivotal location in question, also include the influence of the known boundary values at that same time. That is, all φ values for a given time t_i should be

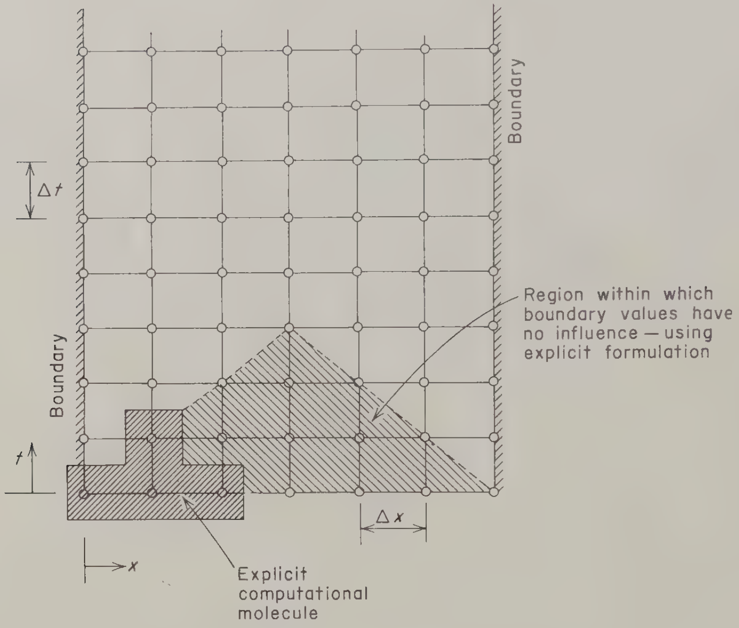

Boundary
Boundary
Δt
t
x
Δx
Region within which boundary values have no influence — using explicit formulation
Explicit computational molecule

Figure 11.92

interrelated—including the boundary values. Such an interrelationship is achieved by using an *implicit-recurrence formula.*

An implicit-recurrence formula is one in which two or more unknown values of the function at t_{i+1} are expressed in terms of known previous values, by a single application of the formula. Such formulations, then, require solution of sets of simultaneous equations.

Implicit-recurrence formulas for parabolic systems can be most readily derived by presuming that variation in the x direction—in the finite-difference formulation of the given differential equation—can be represented by weighted averages of formulations at both t_i and t_{i+1} (see Fig. 11.93). If it is assumed that the contribution at time t_{i+1} is θ, where $0 \leq \theta \leq 1$, the multiplying factor at t_i must be $(1 - \theta)$. For the parabolic equation

$$\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} \tag{11.151}$$

the implicit-recurrence equation is

$$\frac{1}{\Delta t}(-\varphi_0 + \varphi_2) = (1 - \theta)\left(\frac{1}{\Delta x}\right)^2 (\varphi_3 - 2\varphi_0 + \varphi_1) + \theta\left(\frac{1}{\Delta x}\right)^2 (\varphi_5 - 2\varphi_2 + \varphi_4) \tag{11.152}$$

or

$$\varphi_3[r(1-\theta)] + \varphi_0[1 - 2r(1-\theta)] + \varphi_1[r(1-\theta)]$$
$$= \varphi_5(-r\theta) + \varphi_2(1 + 2r\theta) + \varphi_4(-r\theta) \qquad (11.153)$$

where $r = \Delta t/(\Delta x)^2$, and $0 \le \theta \le 1$. The computational molecule associated with Eq. (11.153) is shown in Fig. 11.94. Also illustrated are several special cases, including the previously defined explicit-recurrence formula. Implicit-recurrence formulas are stable for *all* values of r providing $\theta \ge \frac{1}{2}$. The particular case where $\theta = \frac{1}{2}$, that is, the Crank-Nicolson formulation, is most frequently used.

It should be recognized that formulation using the implicit-recurrence computational molecules of Fig. 11.94 results in solution of tridiagonal sets of simultaneous equations. These equations are usually well conditioned and readily solved by any of the methods described in Chap. 5.

Example 11.13 To illustrate the general method of solution, consider again Example 11.9. The equation to be solved is

$$\frac{\partial\varphi}{\partial t} = \frac{\partial^2\varphi}{\partial x^2} \qquad (11.154)$$

and the boundary conditions are specified to be

$$\varphi(0,t) = 0 \qquad \varphi(x,0) = 1.0 \qquad \varphi_x(1,t) = 0 \qquad (11.155)$$

The following particular case will be examined first:

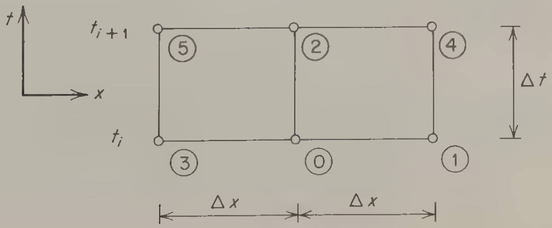
$$\theta = 1.0 \qquad \Delta x = 0.25 \qquad \Delta t = 0.05 \qquad (11.156)$$

Therefore

$$r = \frac{\Delta t}{(\Delta x)^2} = 0.8 \qquad (11.157)$$

The computational molecule is shown in Fig. 11.95, where it has been presumed that the t coordinate increases vertically downward. The notation that will be used in formulating the equations is that defined in Fig. 11.96. For the computational molecule placed in the fashion of

Figure 11.93



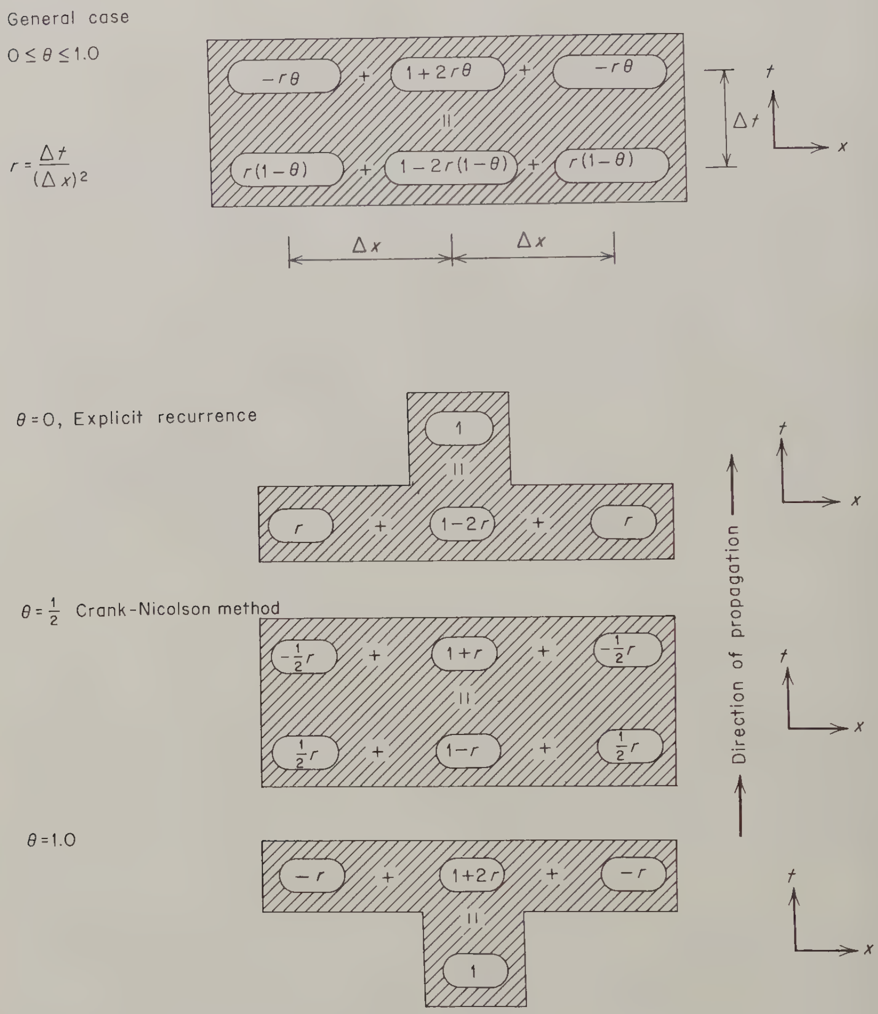
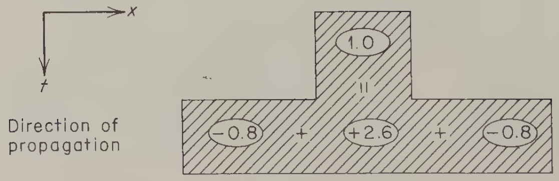


Figure 11.94

Figure 11.95



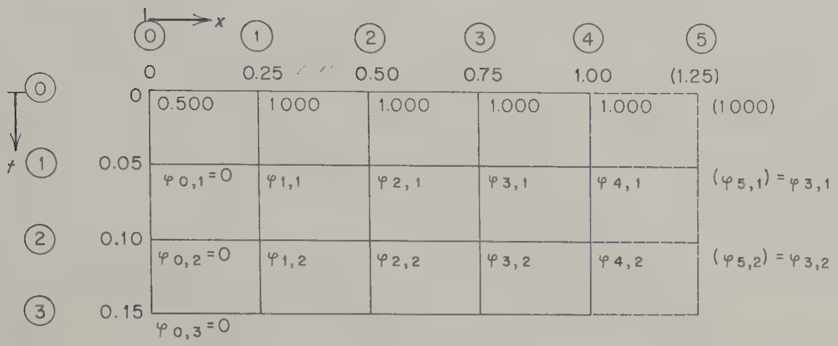


Figure 11.96

Fig. 11.97, the resulting equation is

$$-0.8\varphi_{0,1} + 2.6\varphi_{1,1} - 0.8\varphi_{2,1} = 1.000 \tag{11.158}$$

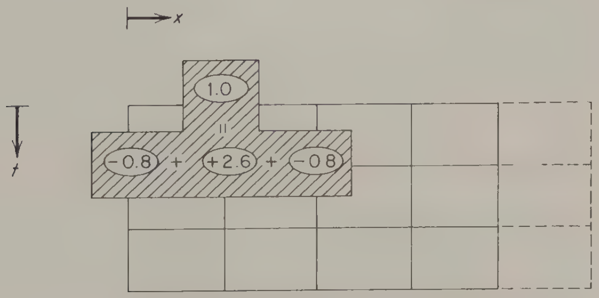
For evaluation at $t = 0.05$, the following equations must be solved:

$$\begin{bmatrix} -0.8 & 2.6 & -0.8 & & & \\ & -0.8 & 2.6 & -0.8 & & \\ & & -0.8 & 2.6 & -0.8 & \\ & & & -0.8 & 2.6 & -0.8 \end{bmatrix} \begin{bmatrix} \varphi_{0,1} \\ \varphi_{1,1} \\ \varphi_{2,1} \\ \varphi_{3,1} \\ \varphi_{4,1} \\ \varphi_{5,1} \end{bmatrix} = \begin{bmatrix} 1.000 \\ 1.000 \\ 1.000 \\ 1.000 \end{bmatrix} \tag{11.159}$$

By noting that $\varphi_{0,1} = 0$ and $\varphi_{5,1} = \varphi_{3,1}$, these equations can be written in the more normal form

$$\begin{bmatrix} 2.6 & -0.8 & & \\ -0.8 & 2.6 & -0.8 & \\ & -0.8 & 2.6 & -0.8 \\ & & -1.6 & 2.6 \end{bmatrix} \begin{bmatrix} \varphi_{1,1} \\ \varphi_{2,1} \\ \varphi_{3,1} \\ \varphi_{4,1} \end{bmatrix} = \begin{bmatrix} 1.000 \\ 1.000 \\ 1.000 \\ 1.000 \end{bmatrix} \tag{11.160}$$

Figure 11.97



Solution is

$$\begin{aligned} \varphi_{1,1} &= 0.656 \\ \varphi_{2,1} &= 0.881 \\ \varphi_{3,1} &= 0.956 \\ \varphi_{4,1} &= 0.973 \end{aligned} \tag{11.161}$$

With these values, solution at $t = 2(0.05) = 0.10$ can now be obtained. The controlling equations are:

$$\begin{bmatrix} 2.6 & -0.8 & & \\ -0.8 & 2.6 & -0.8 & \\ & -0.8 & 2.6 & -0.8 \\ & & -1.6 & 2.6 \end{bmatrix} \begin{bmatrix} \varphi_{1,2} \\ \varphi_{2,2} \\ \varphi_{3,2} \\ \varphi_{4,2} \end{bmatrix} = \begin{bmatrix} 0.656 \\ 0.881 \\ 0.956 \\ 0.973 \end{bmatrix} \tag{11.162}$$

It is to be noted that the coefficient matrix is the same as that of Eqs. (11.160). This will be true for all values of t. It is to be further recognized that the solutions at any given stage become the constants in the simultaneous equations at the next stage.

Example 11.14 For comparison, again with the assumption that $r = 0.8$, the Crank-Nicolson ($\theta = \frac{1}{2}$) formulation of the first set of simultaneous equations is given by

$$\begin{bmatrix} 1.8 & -0.4 & & \\ -0.4 & 1.8 & -0.4 & \\ & -0.4 & 1.8 & -0.4 \\ & & -0.8 & 1.8 \end{bmatrix} \begin{bmatrix} \varphi_{1,1} \\ \varphi_{2,1} \\ \varphi_{3,1} \\ \varphi_{4,1} \end{bmatrix} = \begin{bmatrix} 0.800 \\ 1.000 \\ 1.000 \\ 1.000 \end{bmatrix} \tag{11.163}$$

Solution is

$$\begin{aligned} \varphi_{1,1} &= 0.648 \\ \varphi_{2,1} &= 0.917 \\ \varphi_{3,1} &= 0.980 \\ \varphi_{4,1} &= 0.991 \end{aligned} \tag{11.164}$$

11.7.3 Method of Lines†

Also known as the Runge-Kutta method, the method of lines presumes representation of derivatives in the x direction by finite-difference approximations, while maintaining continuous integration in the t direction. To illustrate the general procedure, again consider the parabolic partial difference equation

$$\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} \tag{11.165}$$

An explicit formulation, assuming "ordinary" finite-difference approximations in the x direction, is

---

† It is to be noted that the method of lines is a technique commonly used to solve partial differential equations on an indirect electrical analog computer.

$$\frac{\partial \varphi_{i,j}}{\partial t} = \frac{1}{(\Delta x)^2} (\varphi_{i-1,j} - 2\varphi_{i,j} + \varphi_{i+1,j}) \tag{11.166}$$

But for such assumptions, φ along each of the preselected pivotal lines is a function only of t. The partial differential equation can, therefore, be expressed as the ordinary differential equation

$$\frac{d\varphi_i}{dt} + \frac{2}{h^2} (\varphi_i) - \frac{1}{h^2} (\varphi_{i-1} + \varphi_{i+1}) = 0 \tag{11.167}$$

where $$h = \Delta x$$

In essence, then, the method of lines reduces the solution of the continuous system to a consideration of the behavior of sets of simultaneous linear first-order ordinary differential equations.

Implicit formulation using the Crank-Nicolson procedure can also be extended to the solution of problems by the method of lines. For most of these cases, the Runge-Kutta equations of Table 10.5 are used for integration in the t direction.

It is to be remembered that the Runge-Kutta methods of Chap. 10 can be used to solve nonlinear ordinary differential equations. The combined method of lines and Runge-Kutta method, implicit or explicit formulation, can therefore be used to solve nonlinear partial differential equation problems.

11.8 Hyperbolic Partial Differential Equations

The two methods most frequently used in the numerical solution of hyperbolic partial differential equations are (1) the method of characteristics and (2) the method of finite differences. The first of these is uniquely suited to the solution of hyperbolic-type systems.

11.8.1 Method of Characteristics

The general concept of the method of characteristics can be illustrated by considering the behavior of a general, single, quasi-linear, first-order partial differential equation.

$$a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = c \tag{11.168}$$

It should be noted that a, b, and c can be constants or functions of u, x, and/or y, but not of $\partial u/\partial x$ or $\partial u/\partial y$.

It is presumed at the outset that the equation in question corresponds

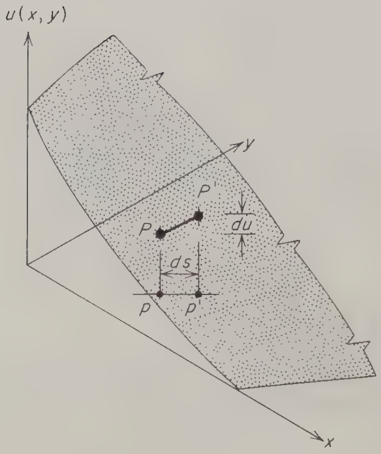


Figure 11.98

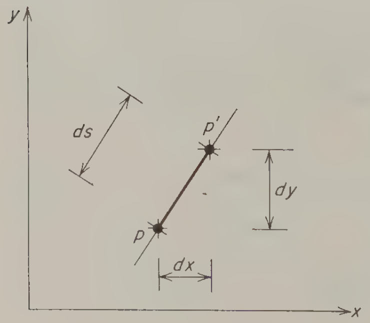
to a single surface in a three-dimensional space whose coordinates are x, y, and $u(x,y)$ (see Figs. 11.98 and 11.99). Moreover, from P to P' on the surface in question, the change in the value of the function u, that is, du, can be related to the rates of change in the function value with respect to the coordinate directions x and y.

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy \tag{11.169}$$

Equations (11.168) and (11.169) constitute a set of two simultaneous equations in the variables $\partial u/\partial x$ and $\partial u/\partial y$.

$$\begin{bmatrix} a & b \\ dx & dy \end{bmatrix} \begin{bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \end{bmatrix} = \begin{bmatrix} c \\ du \end{bmatrix} \tag{11.170}$$

Figure 11.99



If the determinant of the coefficient matrix associated with Eq. (11.170) is not equal to zero, the set will have a unique solution. However, if

$$\begin{vmatrix} a & b \\ dx & dy \end{vmatrix} = 0 \tag{11.171}$$

there will exist an infinite number of possibilities. That is, an infinite number of surfaces could contain the prescribed line PP'.

A curve along which the determinant of the coefficient matrix vanishes is called a *characteristic*. For the single first-order equation considered, the *characteristic direction* is defined by expanding Eq. (11.171).

$$a\,dy = b\,dx$$

or

$$\frac{dy}{b} = \frac{dx}{a} \tag{11.172}$$

The slopes of such lines are

$$\frac{dy}{dx} = \frac{b}{a} \tag{11.173}$$

It is to be noted that these characteristic directions are determined independently of considerations of specific values of $\partial u/\partial x$, $\partial u/\partial y$, and c.

However, no solution to the problem of the determination of $u = u(x,y)$ for a given partial differential equation can be realized unless there are also required the compatible conditions

$$\begin{vmatrix} a & c \\ dx & du \end{vmatrix} = 0 \tag{11.174}$$

and

$$\begin{vmatrix} c & b \\ du & dy \end{vmatrix} = 0 \tag{11.175}$$

These equations result from substitution of the right-hand column of constants for either of the columns of the original coefficient matrix, requiring that the resulting determinant vanish. Expansion of Eqs. (11.174) and (11.175) yields

$$\frac{dx}{a} = \frac{dy}{b} = \frac{du}{c} \tag{11.176}$$

The method of characteristics is concerned primarily with the determination of the particular directions along which integration with respect to only one of the coordinate directions can proceed.

Example 11.15 In the range $0 \le x \le 1.0$, $0 \le y \le 1.0$, solution is desired to the partial differential equation

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 1.0 \tag{11.177}$$

The prescribed initial conditions are

$$u(x,0) = (1.0 - x)^2 \qquad u(0,y) = 1.0 - y^2 \tag{11.178}$$

In terms of the notation of Eq. (11.168)

$$a = b = c = 1.0 \tag{11.179}$$

From Eq. (11.173), the characteristic direction is

$$\frac{dy}{dx} = \frac{b}{a} = 1.0 \tag{11.180}$$

and the equation corresponding to characteristic lines is

$$dy = dx \qquad \text{or} \qquad y = x + (\text{const})_1 \tag{11.181}$$

With Eqs. (11.176), the function value obtained by *integrating along the characteristic* is

$$du = dx \qquad \text{or} \qquad u = x + (\text{const})_2 \tag{11.182}$$

For the specific case illustrated in Fig. 11.100,

$$u - u_0 = x - x_0 \tag{11.183}$$

where $$x_0 = 0.20 \qquad \text{and} \qquad u_0 = 0.64 \tag{11.184}$$

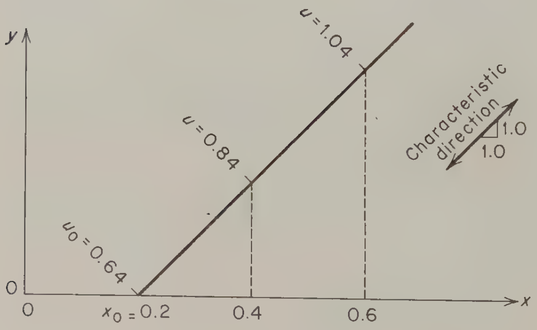
It is to be noted that variation in u along the characteristic is linear. Moreover, only the initial value $u_0(0.20,0) = 0.64$ influences the value of the function along this line.

The complete range of solutions is given in Fig. 11.101.

Consider next the following two partial differential equations:

$$\begin{aligned} a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} &= f \qquad f = f(u,v,x,y) \\ C\frac{\partial v}{\partial x} + D\frac{\partial v}{\partial y} &= F \qquad F = F(u,v,x,y) \end{aligned} \tag{11.185}$$

Figure 11.100



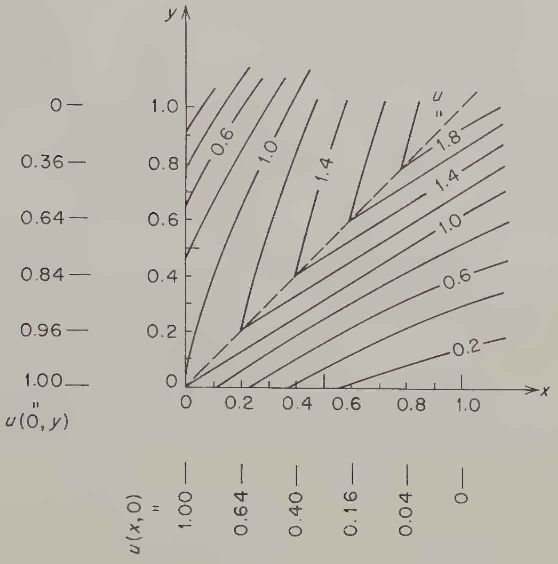
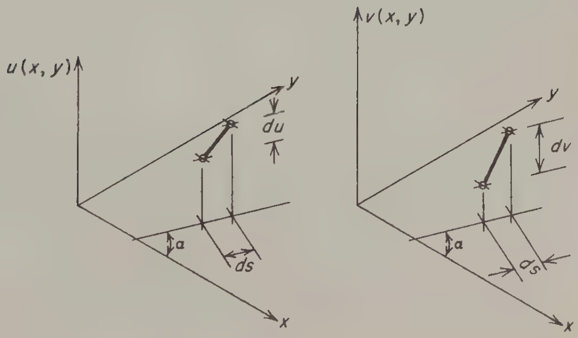


Figure 11.101

It is to be noted that two dependent variables u and v have been presumed. Two surfaces are, therefore, in question (see Fig. 11.102). Two expressions for the differential changes in the function values are, therefore, required. Equations (11.185) and (11.186) constitute two sets of

$$
\begin{aligned}
dx \frac{\partial u}{\partial x} + dy \frac{\partial u}{\partial y} &= du \\
dx \frac{\partial v}{\partial x} + dy \frac{\partial v}{\partial y} &= dv
\end{aligned}
\tag{11.186}
$$

Figure 11.102



two simultaneous equations in the variables $\partial u/\partial x$ and $\partial u/\partial y$, and $\partial v/\partial x$ and $\partial v/\partial y$.

$$\begin{bmatrix} a & b \\ dx & dy \end{bmatrix} \begin{bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \end{bmatrix} = \begin{bmatrix} f \\ du \end{bmatrix} \tag{11.187}$$

and

$$\begin{bmatrix} C & D \\ dx & dy \end{bmatrix} \begin{bmatrix} \dfrac{\partial v}{\partial x} \\ \dfrac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} F \\ dv \end{bmatrix} \tag{11.188}$$

Each of these sets of equations will define a characteristic direction. Each will result in an equation for integration along that characteristic. For Eqs. (11.187)

$$\begin{vmatrix} a & b \\ dx & dy \end{vmatrix} = \begin{vmatrix} a & f \\ dx & du \end{vmatrix} = 0 \tag{11.189}$$

or

$$\frac{dx}{a} = \frac{dy}{b} = \frac{du}{f} \tag{11.190}$$

For Eqs. (11.188)

$$\begin{vmatrix} C & D \\ dx & dy \end{vmatrix} = \begin{vmatrix} C & F \\ dx & dv \end{vmatrix} = 0 \tag{11.191}$$

or

$$\frac{dx}{D} = \frac{dy}{C} = \frac{dv}{F} \tag{11.192}$$

Example 11.16 The equations to be examined are

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 1.0 \qquad \frac{\partial v}{\partial x} - \frac{\partial v}{\partial y} = 1.0 \tag{11.193}$$

The initial conditions, which must be specified for both u and v, are presumed given only for the case $y = 0$.

$$u(x,0) = 0.50 \qquad v(x,0) = 1.00 - x$$

From Eqs. (11.190) it is determined that the direction associated with one of the families of characteristics, say the α characteristic, has a slope with the x axis of $+1$.

$$dx = dy = du \tag{11.194}$$

or

$$\left.\frac{dy}{dx}\right|_{\alpha} = +1 \tag{11.195}$$

Integration along that direction requires that

$$u\Big|_{\alpha} = x + \text{const} \tag{11.196}$$

Using Eqs. (11.192),

$$\frac{dy}{dx}\Big|_{\beta} = -1 \tag{11.197}$$

and

$$-dx = dy = dv \tag{11.198}$$

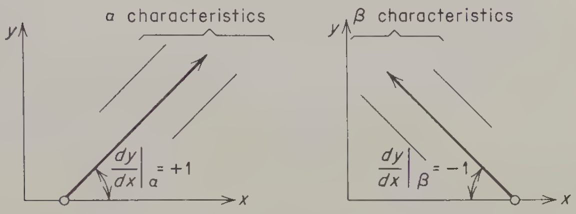
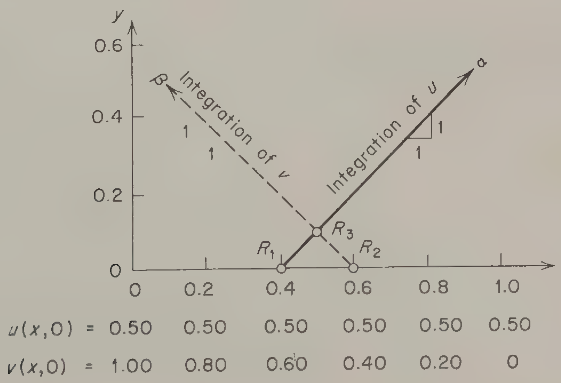


Figure 11.103

The two characteristics α and β can be combined to determine the values of u and v at all their various points of intersection. This is shown for one specific case in Fig. 11.104. Also listed in that figure are initial values of u and v at $\Delta x = 0.2$ intervals.

$$\begin{aligned}&\text{Along } \alpha \text{ characteristic:}\\ &u - u_0 = y - y_0 \quad \text{or} \quad u - 0.50 = 0.10 \quad \text{or} \quad u = 0.60\\ &\text{Along } \beta \text{ characteristic:}\\ &v - v_0 = y - y_0 \quad \text{or} \quad v - 0.40 = 0.10 \quad \text{or} \quad v = 0.50\end{aligned} \tag{11.199}$$

Figure 11.104


y
0.6
0.4
0.2
0
β
Integration of v
1
1
α
Integration of u
1
1
R1
R3
R2
0
0.2
0.4
0.6
0.8
1.0
u(x,0) = 0.50 0.50 0.50 0.50 0.50 0.50
v(x,0) = 1.00 0.80 0.60 0.40 0.20 0

The more general set of two first-order partial differential equations is

$$\begin{aligned} a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + c\frac{\partial v}{\partial x} + d\frac{\partial v}{\partial y} &= f \\ A\frac{\partial u}{\partial x} + B\frac{\partial u}{\partial y} + C\frac{\partial v}{\partial x} + D\frac{\partial v}{\partial y} &= F \end{aligned} \tag{11.200}$$

where a, b, c, d, f, A, B, C, D, and F can be functions of u, v, x, or y, or constants. The full 4 × 4 set of simultaneous equations to be examined is given as Eqs. (11.201).

$$\begin{bmatrix} a & b & c & d \\ A & B & C & D \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{bmatrix} \begin{bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} \\ \dfrac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} f \\ F \\ du \\ dv \end{bmatrix} \tag{11.201}$$

The characteristic directions are obtained by equating the determinant of the coefficient matrix to zero.

$$\begin{vmatrix} a & b & c & d \\ A & B & C & D \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{vmatrix} = 0 \tag{11.202}$$

Expansion yields

$$\left(\frac{dy}{dx}\right)^2 (aC - Ac) - \frac{dy}{dx}(bC - cB + aD - Ad) + (bD - dB) = 0 \tag{11.203}$$

For the determinant

$$\begin{vmatrix} a & f & c & d \\ A & F & C & D \\ dx & du & 0 & 0 \\ 0 & dv & dx & dy \end{vmatrix} = 0 \tag{11.204}$$

$$du\left[aD - Ad + \frac{dy}{dx}(-aC + Ac)\right] + dv(cD - Cd) + dx\left[Fd - fD + \frac{dy}{dx}(fC - Fc)\right] = 0 \tag{11.205}$$

For the case where

$$\begin{vmatrix} a & b & c & f \\ A & B & C & F \\ dx & dy & 0 & du \\ 0 & 0 & dx & dv \end{vmatrix} = 0 \tag{11.206}$$

$$du(-aB + Ab) + dv\left[bC - Bc + \frac{dy}{dx}(-aC + Ac)\right] + \left[Bf - bF + \frac{dy}{dx}(Fa - Af)\right] = 0 \tag{11.207}$$

It is to be recognized that for hyperbolic systems Eq. (11.203) defines two characteristic directions. Integration of either of Eqs. (11.205) or (11.206) in these two directions will result in the desired solution.

Example 11.17

$$\begin{bmatrix} 1 & 0 & 0 & \eta \\ 0 & \eta & 1 & 0 \\ dx & dy & 0 & 0 \\ 0 & 0 & dx & dy \end{bmatrix} \begin{bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} \\ \dfrac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ du \\ dv \end{bmatrix} \tag{11.208}$$

From Eq. (11.203), substituting the corresponding values for $a, b, \ldots, f$ and $A, B, \ldots, F$ from Eq. (11.208),

$$\left(\frac{dy}{dx}\right)^2 (1) - \frac{dy}{dx}(0) + (-\eta^2) = 0$$

or

$$\frac{dy}{dx} = \pm\sqrt{\eta^2} = \pm\eta \tag{11.209}$$

Similar substitution into Eq. (11.205) yields

$$du\left(-\frac{dy}{dx}\right) + dv(\eta) = 0 \tag{11.210}$$

The total solution is, therefore, that summarized in Table 11.13.

Along a boundary (see Fig. 11.105), the set of initial conditions given in Table 11.14 is prescribed for the set of equations listed as (11.208). $\eta = 2.0$ is also specified. To determine the values of u and v at location

Table 11.13

α Characteristic	β Characteristic
$\left(\frac{dy}{dx}\right)_\alpha = +\eta$	$\left(\frac{dy}{dx}\right)_\beta = -\eta$
$du(\eta) + dv(\eta) = 0$ or $u + v = c_1$†	$du(-\eta) + dv(\eta) = 0$ or $u - v = c_2$†

† c_1 and c_2 are constants, depending on the initial conditions.

$P_{4,6}$ (see Fig. 11.105),

$$\begin{aligned} u + v &= c_1 = 1.60 \qquad \text{(from point ④)} \\ u - v &= c_2 = -1.40 \qquad \text{(from point ⑥)} \end{aligned} \tag{11.211}$$

or $$u = \tfrac{1}{2}(c_1 + c_2) = 0.10 \qquad v = \tfrac{1}{2}(c_1 - c_2) = 1.50 \tag{11.212}$$

Solutions at all possible intersection points are listed in Table 11.15. The points of intersection are indicated in Fig. 11.106.

Consider, now, the solution of a second-order quasi-linear hyperbolic partial differential equation. In its most general form

$$a\frac{\partial^2\varphi}{\partial x^2} + b\frac{\partial^2\varphi}{\partial x\,\partial y} + c\frac{\partial^2\varphi}{\partial y^2} = f \tag{11.213}$$

where a, b, c, and f can be functions of x, y, φ, $\partial\varphi/\partial x$, and $\partial\varphi/\partial y$, or constants. To facilitate representation, the following standard notation

Figure 11.105

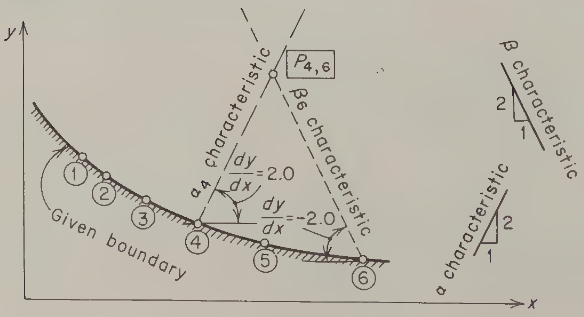


Table 11.14

Location:	①	②	③	④	⑤	⑥
u	1.00	1.00	0.90	0.80	0.70	0.60
v	0.10	0.20	0.40	0.80	1.20	2.00
$c_1 = u + v$	1.10	1.20	1.30	1.60	1.90	2.60
$c_2 = u - v$	0.90	0.80	0.50	0	-0.50	-1.40

Table 11.15

β Characteristics	α Characteristics					
	α_1	α_2	α_3	α_4	α_5	α_6
	$c_1 = 1.10$	$c_1 = 1.20$	$c_1 = 1.30$	$c_1 = 1.60$	$c_1 = 1.90$	$c_1 = 2.60$
β_1 $c_2 = 0.90$						
β_2 $c_2 = 0.80$	$u = 0.95$ $v = 0.15$					
β_3 $c_2 = 0.50$	$u = 0.80$ $v = 0.30$	$u = 0.85$ $v = 0.35$				
β_4 $c_2 = 0$	$u = 0.55$ $v = 0.55$	$u = 0.60$ $v = 0.60$	$u = 0.65$ $v = 0.65$			
β_5 $c_2 = -0.50$	$u = 0.30$ $v = 0.80$	$u = 0.35$ $v = 0.85$	$u = 0.40$ $v = 0.90$	$u = 0.55$ $v = 1.05$		
β_6 $c_2 = -1.40$	$u = -0.15$ $v = 1.25$	$u = -0.10$ $v = 1.30$	$u = -0.05$ $v = 1.35$	$u = 0.10$ $v = 1.50$	$u = 0.25$ $v = 1.65$	

will be presumed:

$$
\begin{aligned}
p &= \frac{\partial \varphi}{dx} \\
q &= \frac{\partial \varphi}{\partial y} \\
r &= \frac{\partial^2 \varphi}{\partial x^2} \\
s &= \frac{\partial^2 \varphi}{\partial x \, \partial y} \\
t &= \frac{\partial^2 y}{\partial y^2}
\end{aligned}
\tag{11.214}
$$

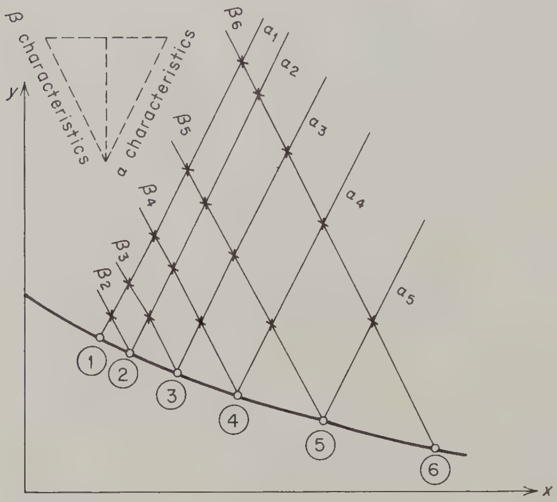


Figure 11.106

In terms of these quantities, the required differentials are

$$du = p\,dx + q\,dy \tag{11.215}$$

$$dp = r\,dx + s\,dy \tag{11.216}$$

$$dq = s\,dx + t\,dy \tag{11.217}$$

and the original differential equation is

$$ar + bs + ct = f \tag{11.218}$$

Equations (11.216) to (11.218) can be written in matrix form in the variables r, s, and t.

$$\begin{bmatrix} a & b & c \\ dx & dy & 0 \\ 0 & dx & dy \end{bmatrix} \begin{bmatrix} r \\ s \\ t \end{bmatrix} = \begin{bmatrix} f \\ dp \\ dq \end{bmatrix} \tag{11.219}$$

By proceeding as before, the characteristic direction is determined by equating the determinant of the coefficient matrix to zero.

$$\begin{vmatrix} a & b & c \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = 0 \tag{11.220}$$

or

$$a\left(\frac{dy}{dx}\right)^2 - b\,\frac{dy}{dx} + c = 0 \tag{11.221}$$

The equation is hyperbolic if $b^2 - 4ac > 0$. That is, Eq. (11.221) will have associated with it two "real" solutions for dy/dx.

Substituting the right-hand column of constants for the last column of the coefficient matrix and again equating the resulting determinant to zero yields

$$\begin{vmatrix} a & b & f \\ dx & dy & dp \\ 0 & dx & dq \end{vmatrix} = 0 \tag{11.222}$$

or

$$\left(b - a\frac{dy}{dx}\right) dq + a\,dp - f\,dx = 0 \tag{11.223}$$

Substitution in the second column yields

$$\begin{vmatrix} a & f & c \\ dx & dp & 0 \\ 0 & dq & dy \end{vmatrix} = 0 \tag{11.224}$$

or

$$c\,dq + a\frac{dy}{dx}\,dp - f\,dy = 0 \tag{11.225}$$

Example 11.18 For the region $0 \leq x \leq 1$, $t > 0$, the simple wave equation

$$\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2} \tag{11.226}$$

is specified. The initial and boundary conditions are

$$\begin{aligned} \varphi(x,0) &= 4\epsilon(1 - x)x \\ \varphi(0,t) &= 0 \\ \varphi(1.0,t) &= 0 \\ q(x,0) = \varphi_t(x,0) &= 0 \end{aligned} \tag{11.227}$$

By recasting the given equation in the form of Eqs. (11.213),

$$\frac{\partial^2\varphi}{\partial x^2} - \frac{\partial^2\varphi}{\partial t^2} = 0 \tag{11.228}$$

the coefficients a, b, c, and f have the following values:

$$a = 1 \qquad b = 0 \qquad c = -1 \qquad f = 0 \tag{11.229}$$

From Eq. (11.221), the characteristic directions are

$$\left(\frac{dt}{dx}\right)^2 (1) - \frac{dt}{dx}(0) - 1 = 0$$

or

$$\frac{dt}{dx} = \pm 1 \tag{11.230}$$

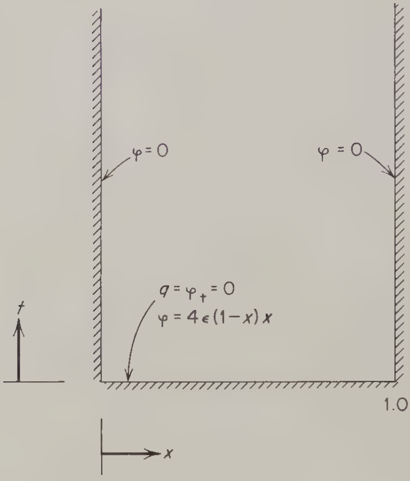


Figure 11.107

Along each of these, by using Eq. (11.225), integration would proceed according to

$$dq(-1) + dp\left[1\left(\frac{dt}{dx}\right)\right] + 0(dt) = 0$$

or

$$dq = \frac{dt}{dx}\,dp \tag{11.231}$$

The solution is summarized in Table 11.16. If an $x = 0.2$ interval for numerical computation is presumed, the initial values at the preselected pivotal location are those listed in Table 11.17. It is to be noted that since φ is specified at $t = 0$, p along this boundary is determined by differentiation.

Table 11.16

α Characteristic	β Characteristic
$\left(\frac{dt}{dx}\right)_\alpha = +1$	$\left(\frac{dt}{dx}\right)_\beta = -1$
$dq(1) - dp(1) = 0$ or $q - p = c_1$	$dq(1) - dp(-1) = 0$ or $q + p = c_2$

Table 11.17

Location	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1.0$
φ	0	0.64ϵ	0.96ϵ	0.96ϵ	0.64ϵ	0
q	0	0	0	0	0	0
p	4.00ϵ	2.40ϵ	0.80ϵ	-0.80ϵ	-2.40ϵ	-4.00ϵ
$q - p = c_1$	-4.00ϵ	-2.40ϵ	-0.80ϵ	$+0.80\epsilon$	$+2.40\epsilon$	$+4.00\epsilon$
$q + p = c_2$	$+4.00\epsilon$	$+2.40\epsilon$	$+0.80\epsilon$	-0.80ϵ	-2.40ϵ	-4.00ϵ

$$p = \frac{\partial \varphi}{\partial x} = \frac{\partial}{\partial x}[4\epsilon(1 - x)x] = 4\epsilon(1 - 2x) \tag{11.232}$$

Similarly, along the boundaries $x = 0$ and $x = 1.0$,

$$q = \frac{\partial \varphi}{\partial t} = \frac{\partial}{\partial t}(0) = 0 \tag{11.233}$$

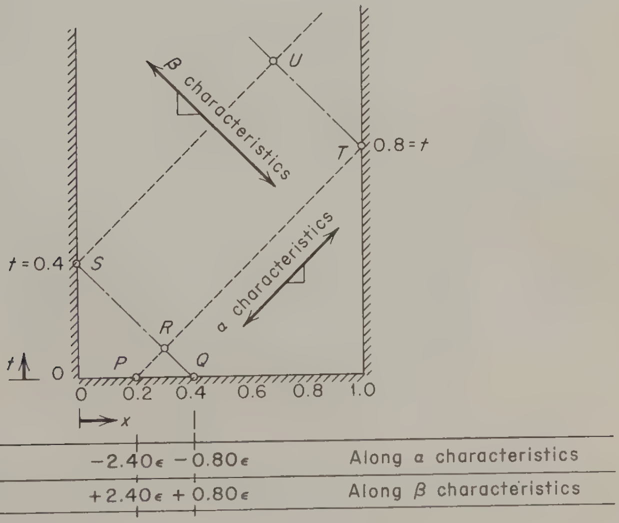
At location P, $q = 0$ and $p = 2.40\epsilon$. At Q, $q = 0$ and $p = 0.80\epsilon$. At location R,

$$\begin{aligned} q - p &= (c_1)_\alpha = -2.40\epsilon \\ q + p &= (c_2)_\beta = +0.80\epsilon \end{aligned} \tag{11.234}$$

This gives

$$\begin{aligned} q &= \tfrac{1}{2}(c_1 + c_2) = -0.80\epsilon \\ p &= \tfrac{1}{2}(c_1 - c_2) = -1.60\epsilon \end{aligned} \tag{11.235}$$

Figure 11.108



At location S, it is known from the β characteristic originating at location Q that

$$q + p = +0.80\epsilon \tag{11.236}$$

But the boundary condition at S, $\varphi(0,t) = 0$, requires that q along $x = 0$ equal zero. Therefore,

$$+p_S = +0.80\epsilon \tag{11.237}$$

This in turn requires that the new α characteristic (from S to U) have the value

$$q - p = 0 - 0.80\epsilon = c_1 \tag{11.238}$$

In a similar manner, the values of q and p associated with location T are

$$\begin{aligned} q &= 0 \\ q - p &= -2.40\epsilon \qquad (11.239) \\ \text{or} \qquad p &= 2.40\epsilon \end{aligned}$$

The β characteristic from T to U is controlled by the equation

$$q + p = 2.40\epsilon = c_2 \tag{11.240}$$

At location U,

$$\left.\begin{aligned} q &= \tfrac{1}{2}(c_{1\alpha} + c_{2\beta}) = \tfrac{1}{2}(-0.80\epsilon + 2.40\epsilon) = 0.80\epsilon \\ p &= \tfrac{1}{2}(c_{1\alpha} - c_{2\beta}) = \tfrac{1}{2}(-0.80\epsilon - 2.40\epsilon) = -1.60\epsilon \end{aligned}\right. \tag{11.241}$$

It is to be recognized that φ values at the various points of intersection are obtained by analytical or numerical integration of q (that is, $\partial\varphi/\partial t$) or p (that is, $\partial\varphi/\partial x$) values, starting from the boundaries.

Example 11.19 In all the preceding examples of this section, linear equations were presumed. Moreover, constants that result in easily integrated functions were selected. The method of characteristics works equally well for nonlinear problems. However, for most of these latter cases, integration is carried out by means of approximate numerical methods. To illustrate a procedure that might be followed, consider a second-order nonlinear partial differential equation of the type defined in Eq. (11.242).

$$a\frac{\partial^2\varphi}{\partial y^2} + b\frac{\partial^2\varphi}{\partial x\,\partial y} + c\frac{\partial^2\varphi}{\partial y^2} = f \tag{11.242}$$

In general, the equations for the characteristic directions will vary with x and/or y. Curves (rather than straight lines) will normally result.

Three equations are of importance:

$$a\left(\frac{dy}{dx}\right)^2 - b\,\frac{dy}{dx} + c = 0 \tag{11.243}$$

$$c\,dq + a\,\frac{dy}{dx}\,dp - f\,dy = 0 \tag{11.244}$$

$$du = p\,dx + q\,dy \tag{11.245}$$

If it is presumed that Eq. (11.243) has two real roots α and β associated with each set of values of x and y, the values of x_R, y_R, q_R, p_R, and φ_R associated with location R (see Fig. 11.109) can be determined from the following set of five nonlinear algebraic equations:

$$y_R - y_P = [\tfrac{1}{2}(\alpha_R + \alpha_P)](x_R - x_P) \tag{11.246}$$

$$y_R - y_Q = [\tfrac{1}{2}(\beta_R + \beta_Q)](x_R - x_Q) \tag{11.247}$$

$$(q_R - q_P)[\tfrac{1}{2}(c_R + c_P)] + (p_R - p_P)[\tfrac{1}{2}(a_R\alpha_R + a_P\alpha_P)] - (y_R - y_P)[\tfrac{1}{2}(f_R + f_P)] = 0 \tag{11.248}$$

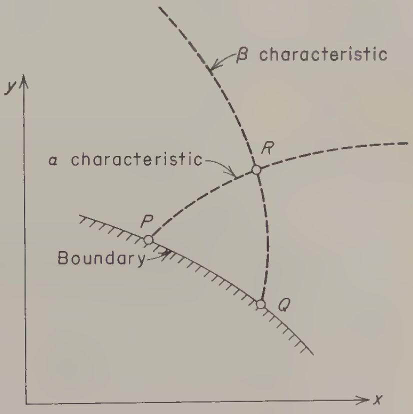
$$(q_R - q_\varphi)[\tfrac{1}{2}(c_R + c_P)] + (p_R - p_Q)[\tfrac{1}{2}(a_R\beta_R + a_Q\beta_Q)] - (y_R - y_Q)[\tfrac{1}{2}(f_R + f_Q)] = 0 \tag{11.249}$$

$$u_R - u_P = [\tfrac{1}{2}(p_R + p_P)](x_R - x_P) + [\tfrac{1}{2}(q_R + q_P)](y_R - y_P) = 0 \tag{11.250}$$

The solution procedure is normally an iterative one.

Two factors concerning initial values and the use of the method of characteristics in the solution of hyperbolic systems are to be noted:

Figure 11.109



First, discontinuous initial values will propagate into the solution domain along characteristic lines. No undue complication is introduced by such a condition. Secondly, when the boundary along which initial conditions are specified coincides with a characteristic curve for the equation(s) in question, the values of the initial conditions must be identical to the integrated solution along the characteristics.

11.8.2 Finite-difference Method

While the method of characteristics is the most accurate and general of the several methods for solving hyperbolic partial differential equations, the finite-difference method can be used to solve many of the problems that arise in practice. Again, as demonstrated in Sec. 11.7 on solution of parabolic equations, two finite-difference formulations are possible: explicit and implicit. Again, pivotal spacing must be selected to ensure stability.

EXPLICIT FORMULATION

For the particular hyperbolic equation (the one-dimensional wave equation)

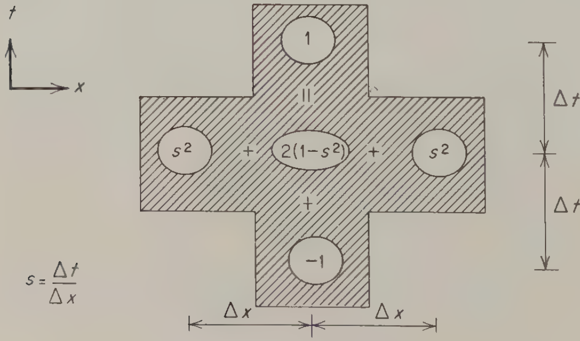
$$\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} = 0 \tag{11.251}$$

using ordinary second-order central differences (see Table 9.27), the computational molecule is that shown in Fig. 11.110. For convergence

$$s = \frac{\Delta t}{\Delta x} \leq 1.0 \tag{11.252}$$

For optimum accuracy, the selected grid coordinate ratios, that is, $\Delta t/\Delta x$, should be such that the region of finite-difference solution coincides

Figure 11.110



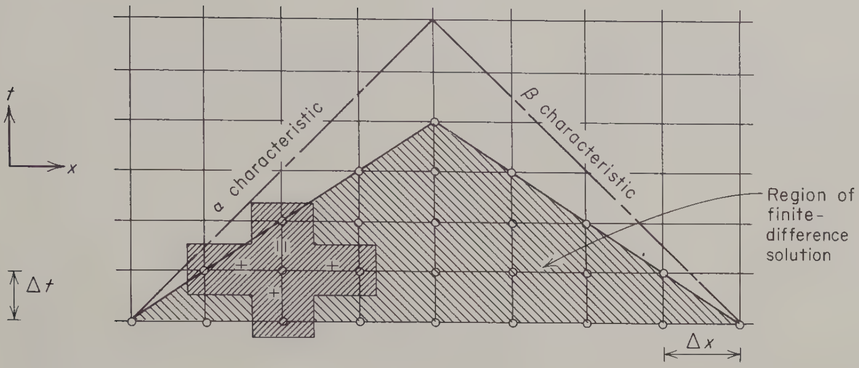


Figure 11.111

with the region of solution defined by the α and β characteristics of the given equation (see Fig. 11.111).

Example 11.20 The given equation is

$$\frac{\partial^2 \varphi}{\partial t^2} - \frac{\partial^2 \varphi}{\partial x^2} = 0 \tag{11.253}$$

The initial and boundary conditions are

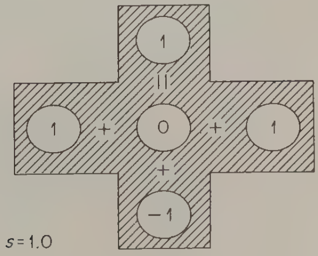
$$\begin{aligned} \varphi(x,0) &= 4(1-x)(x) \\ \varphi(0,t) &= 0 \\ \varphi(1.0,t) &= 0 \\ \varphi_t(x,t) &= 0 \end{aligned} \tag{11.254}$$

The region in question is

$$0 \le x \le 1 \qquad \text{and} \qquad t > 0 \tag{11.255}$$

If a square grid is presumed, that is, $s = 1.0$, the computational molecule for continuing a solution is that shown in Fig. 11.112. To handle the

Figure 11.112



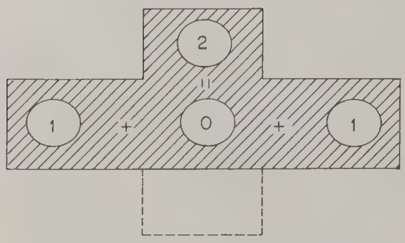


Figure 11.113

special problem associated with obtaining the first set of solutions at $t = \Delta t$, it is to be recognized that the last of the four conditions specified in Eqs. (11.254) requires that

$$\varphi(x,\Delta t) = \varphi(x,-\Delta t) \tag{11.256}$$

The computational molecule for that special "first" case is shown in Fig. 11.113. The solution through one full period is given in Fig. 11.114, presuming that $\Delta x = 0.2$.

Figure 11.114

t \ x	0	0.2	0.4	0.6	0.8	1.0
	0	+0.48	+0.80	+0.80	+0.48	0
	0	+0.64	+0.96	+0.96	+0.64	0
	0	+0.48	+0.80	+0.80	+0.48	0
	0	+0.16	+0.32	+0.32	+0.16	0
	0	−0.16	−0.32	−0.32	−0.16	0
	0	−0.48	−0.80	−0.80	−0.48	0
	0	−0.64	−0.96	−0.96	−0.64	0
	0	−0.48	−0.80	−0.80	−0.48	0
0.6	0	−0.16	−0.32	−0.32	−0.16	0
0.4	0	+0.16	+0.32	+0.32	+0.16	0
0.2	0	+0.48	+0.80	+0.80	+0.48	0
0	0	+0.64	+0.96	+0.96	+0.64	0

IMPLICIT FORMULATION

If it is assumed that the finite-difference expression for $\partial^2\varphi/\partial x^2$ at location $t = t_i$ can be represented by the average of the second-order finite differences at $t = t_i + \Delta t$ and $t = t_i - \Delta t$, the Crank-Nicolson computational molecule for the equation

$$\frac{\partial^2\varphi}{\partial t^2} - \frac{\partial^2\varphi}{\partial x^2} = 0 \tag{11.257}$$

is that shown in Fig. 11.115. Solution using this formulation will converge, independent of the value of s.

Prob. 11.1: Develop higher-order computational molecules for ∇^2 and ∇^4 in the following coordinate systems:

(*a*) 30° skewed (*b*) 45° skewed
(*c*) Triangular ($\alpha = 30°$, $\beta = 60°$) (*d*) Polar

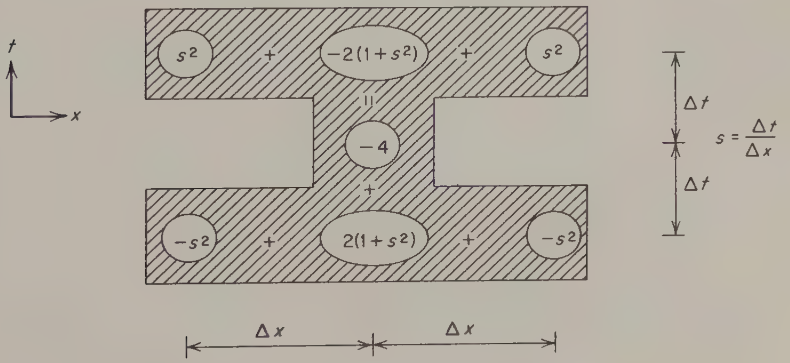
Prob. 11.2: For the square region $L \times L$, it is specified that values of φ on the boundary are those illustrated in Fig. 11.116. If $\nabla^2\varphi = 0$ within the region, determine the values of φ at the pivotal locations, presuming grid spacings equal to

(*a*) $\frac{1}{2}L$ (*b*) $\frac{1}{3}L$ (*c*) $\frac{1}{4}L$ (*d*) $\frac{1}{5}L$

Prob. 11.3: If, for Fig. 11.116, derivatives of φ normal to each edge are also required to be equal to zero, and if $\nabla^4\varphi = K$ is specified for the interior of the region, determine the values of φ at the pivotal locations, presuming grid spacings equal to

(*a*) $\frac{1}{2}L$ (*b*) $\frac{1}{3}L$ (*c*) $\frac{1}{4}L$ (*d*) $\frac{1}{5}L$

Figure 11.115



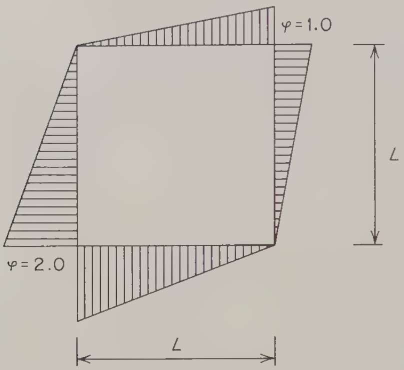


Figure 11.116

Prob. 11.4: Determine the values of Ψ at the five pivotal locations indicated in Fig. 11.117. The interior of the region is controlled by the equation

$$\nabla^2\Psi = \Psi$$

Ψ on all external boundaries is equal to zero.

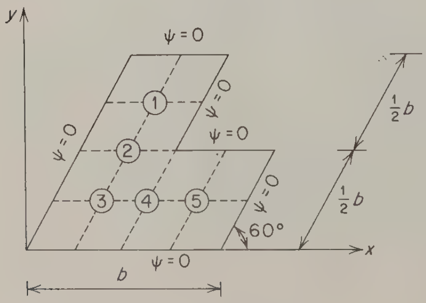
Prob. 11.5: Determine the first three eigenvalues for the circular plate with hole, whose edges are simply supported, as shown in Fig. 11.118. The controlling equation is

$$\nabla^2 w + \lambda w = 0$$

where w is the lateral deflection of the plate.

Prob. 11.6: A triangular plate (plan view shown in Fig. 11.119) is fixed along its longer edge, that is, $w = 0$ and $w_\eta = 0$, and simply supported along each of its equal-leg shorter edges, that is, $w = 0$ and

Figure 11.117



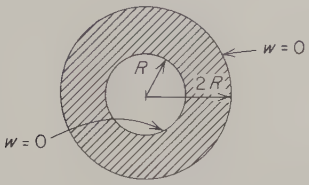


Figure 11.118

$w_{\eta\eta} = 0$. The equation governing lateral deflection of the plate is

$$\nabla^4 w = \frac{q}{D}$$

where q is lateral load per unit area, and D is the plate flexural rigidity constant. Determine the maximum deflection of the plate, solving the problem by presuming three different grid spacings and extrapolating these results to an infinite number of pivotal locations.

Prob. 11.7: Develop group-relaxation operators similar to those shown in Figs. 11.48 to 11.50 and 11.52 for 30, 45, and 60° skewed-coordinate systems.

Prob. 11.8: Develop computational molecules similar to those listed in Table 11.10 for unequal grids, presuming patterns having spatial dimensions related by factors of 3.

Prob. 11.9: For the region and grid spacing defined by Fig. 11.120, given the equation

$$\nabla^4 w + \lambda \frac{\partial^2 w}{\partial x^2} = 0$$

the boundary conditions are

$$\begin{aligned} w(0,y) &= 0 \\ w(1.0,y) &= 0 \\ w(x,0) &= 0 \\ w_y(x,0) &= 0 \\ w_{xx}(0,y) &= 0 \\ w_{xx}(1.0,y) &= 0 \\ w_{yy}(x,1.0) &= 0 \\ w_{xxx}(x,1.0) + 2w_{xyy}(x,1.0) &= 0 \end{aligned}$$

Determine the first three eigenvalues associated with the system.

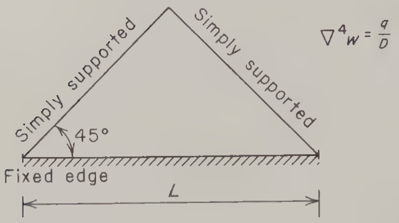


Figure 11.119

Prob. 11.10: Discuss the numerical stability of the explicit finite-difference formulation of the equation

$$\frac{\partial z}{\partial t} = \nabla^2 z \qquad z = z(x,y,t)$$

presuming ordinary second-order finite differences, $\Delta x = \Delta y$, and

$$\frac{\Delta t}{(\Delta x)^2} = \tfrac{1}{2} \text{ and } 1.0$$

Prob. 11.11: The vibration of a beam is governed by the equation

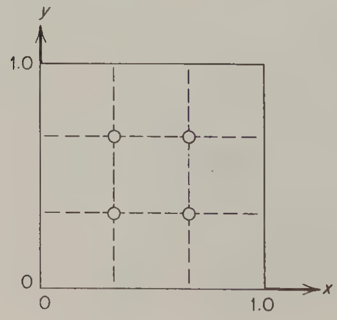
$$\frac{\partial^4 w}{\partial x^4} + \frac{\partial^2 w}{\partial t^2} = 0$$

For the initial and boundary conditions

$$w(0,t) = 0$$
$$w(1.0,t) = 0$$
$$w_x(0,t) = 0$$
$$w_x(1.0,t) = 1.0$$

define the behavior of the beam, through at least one period of vibration.

Figure 11.120



Prob. 11.12: By both the method of characteristics and the method of finite differences, solve the following four partial differential equations in the range $0 \leq x \leq 1.0$, $0 \leq y \leq 1.0$:

$$\frac{\partial \varphi}{\partial x} + \frac{\partial u}{\partial y} = 0$$

$$\frac{\partial \Psi}{\partial x} + \frac{\partial v}{\partial y} = 0$$

$$\frac{\partial u}{\partial y} = A = 1.0$$

$$\frac{\partial v}{\partial y} = B = 1.0$$

The initial and boundary conditions are

$$\varphi(0,y) = 1.0$$

$$\Psi(0,y) = 0$$

$$u(x,0) = 0$$

$$v(x,0) = 0$$

Prob. 11.13: Show that Eqs. (11.190) and (11.192) are special cases of Eqs. (11.203), (11.205), and (11.207).

Prob. 11.14:

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = 0 \qquad u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = 0$$

The initial conditions are

$$v(x,0) = (1 - x)^2$$

$$u(0 \leq x \leq 0.5,\ 0) = 1.0$$

$$u(0.5 \leq x \leq 1.0,\ 0) = x$$

Determine $u(x,y)$ and $v(x,y)$ in the range $0 \leq y \leq 1.0$.

Prob. 11.15:

$$\begin{bmatrix} a & 1 \\ 1 & L \end{bmatrix} \begin{bmatrix} \dfrac{\partial \varphi}{\partial t} \\ \dfrac{\partial u}{\partial x} \end{bmatrix} = \begin{bmatrix} -\varphi \\ u \end{bmatrix}$$

For the special case where $a = 1$ and $L = 4$, determine the values of φ and u in the range $0 \leq x \leq 1.0$, and $0 \leq t \leq 2.0$. The initial conditions are

$$\varphi(0,0) = 1.0$$

$$\varphi(x \neq 0, 0) = 0$$

$$u(x,t) = 0$$

Prob. 11.16:

$$\frac{\partial u}{\partial x} + v = 0$$

$$\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = 0$$

$$u(0,y) = 0$$

$$v(0,y) = 1.0 - x$$

Determine u and v in the range $0 \leq x \leq 1.0$, $0 \leq y \leq 1.0$.

optimization

12

In its most elementary and basic form, *optimization* is concerned with the determination of the "best" solution to a given problem. This process is required in the solution of many problems in engineering and applied science—in the maximization (or minimization) of a given function(s), in the selection of *control variables* to facilitate realization of a desired condition, in the scheduling of a series of operations or events to control completion dates of a given project, in the development of *optimal* layouts of organizational units within a given design space, etc. It is not the intent of this chapter to develop all or even a majority of the optimizing methods that have been proposed, but rather to describe several of the more often referred to systems.

12.1 Introduction: Concepts and Definitions

In general, the methods developed in this section are extensions of the theory of ordinary maxima and minima of functions of one variable developed in elementary calculus. It is to be recalled that in those cases a function $f(x)$ is said to have a maximum value at $x = a$ if

$$\Delta f^+ \equiv f(a+h) - f(a) < 0$$

and

$$\Delta f^- \equiv f(a-h) - f(a) < 0 \tag{12.1}$$

for small values of h. For the case where Δf^+ and Δf^- are both positive, a minimum is realized.

By assuming that $f(x)$ is continuous in the interval in question, and moreover presuming that it possesses continuous derivatives through order n, Eqs. (12.1) can be expanded in Taylor series as indicated in Eqs. (12.2).

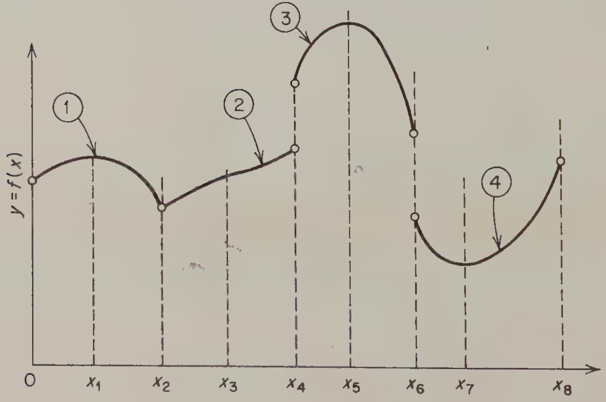
$$\begin{aligned} \Delta f^+ &= f'(a)h + \frac{1}{2!}f''(a)h^2 + \frac{1}{3!}f'''(a)h^3 + \cdots \lessgtr 0 \\ \Delta f^- &= -f'(a)h + \frac{1}{2!}f''(a)h^2 - \frac{1}{3!}f'''(a)h^3 + \cdots \lessgtr 0 \end{aligned} \tag{12.2}$$

The necessary condition for a maximum is that the first nonvanishing (i.e., nonzero) derivative at $x = a$ be of even order and have a negative value. For minimization, again the first nonvanishing derivative must be of even order, but it must have a positive value. For that case where the first nonvanishing derivative is of odd order, and $f''(a) = 0$, a *point of inflection* is said to exist.

It is to be noted that a maximum (or minimum) can be realized without having $f'(x) = 0$. Consider the series of function(s) shown in Fig. 12.1. In the range $0 \leq x \leq x_2$, a continuous function ① is assumed to exist. In the range $x_2 \leq x \leq x_4$, the function has been noted as ②. At x_4 and x_6 a discontinuity in the function is presumed. ③ and ④ are continuous functions in the ranges $x_4 \leq x \leq x_6$ and $x_6 \leq x \leq x_8$, respectively. Consider each of these regions independently:

1. $0 \leq x \leq x_2$: A (local) maximum occurs at x_1, where $f'(x_1) = 0$ and the smallest $f^{\text{even}}(x_1) < 0$. A (local) minimum value of the func-

Figure 12.1



tion in the interval in question occurs at the boundary value x_2, where $f'(x) \neq 0$.

2. $x_2 \leq x \leq x_4$: A (local) minimum occurs at x_2, where $f'(x_2) \neq 0$. A (local) maximum is realized at the boundary value x_4. Here, again, $f'(x_4) \neq 0$. At x_3 there is a point of inflection; that is, $f''(x_3) = 0$.
3. $x_4 \leq x \leq x_6$: A (local) maximum is observed at x_5, where $f'(x_5) = 0$ and the smallest $f^{\text{even}}(x_5) < 0$. A (local) minimum exists at the boundary x_6, with $f'(x_6) \neq 0$.
4. $x_6 \leq x \leq x_8$: $x = x_7$ is the location of the (local) minimum value of the function. Here, $f'(x_7) = 0$ and the earliest $f^{\text{even}}(x_7) \neq 0$.

It is to be noted that in each of these ranges, the *local* (or relative) maximum and minimum values have been noted. The *absolute* maximum and minimum values for the entire range $0 \leq x \leq x_8$ are obtained by comparing the numerical values of these local extremes.

$$f(x)_{\max} = f(x_5) \qquad f(x)_{\min} = f(x_7) \tag{12.3}$$

For many problems in engineering and applied science it is possible to establish at the outset that for a particular interval in question, the function possesses one hump (or one depression). This property is known as *unimodality.* In mathematical terms, unimodality requires the following:†

$$\begin{aligned} &\text{For } x_a < x_b < x^*: && y_a < y_b < y^* \\ &\text{For } x^* < x_a < x_b: && y^* > y_a > y_b \end{aligned} \tag{12.4}$$

In Fig. 12.1 the described function is unimodal in the interval $x_2 \leq x \leq x_6$.

The concept of a *convex* function is more restrictive than that of unimodality. If three values of the independent variable $x_a < x_b < x_c$ are considered, convexity requires that the true value of y_b be greater than that which corresponds to a linear interpolation between y_a and y_c. That is,

$$y_b \geq y_a \frac{x_c - x_b}{x_c - x_a} + y_c \frac{x_b - x_a}{x_c - x_a} \tag{12.5}$$

The type of reasoning used in the above-described single-variable case can be extended to the solution of problems having a larger number of independent parameters. For that particular case where $n = 2$, that is, $f(x_1x_2)$, the behavior can be represented by a contour plot of equal values of the dependent function f. Consider, for example, that particular case described in Fig. 12.2. The function is said to have a maxi-

---

† x^* is the value of x corresponding to the maximum value of $y = y^*$.

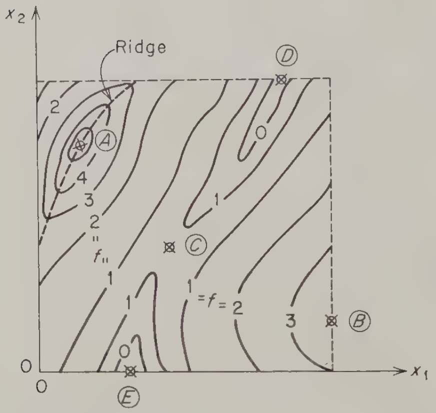


Figure 12.2

mum at $x_1 = a$ and $x_2 = b$ if

$$f(a \pm h, b \pm k) - f(a,b) < 0 \tag{12.6a}$$

and a minimum if

$$f(a \pm h, b \pm k) - f(a,b) > 0 \tag{12.6b}$$

for small values of h and k.

The Taylor series expansion of the function in two variables about a point a, b yields

$$f(x_1,x_2) = f(a,b) + f_{x_1}(a,b)h + f_{x_2}(a,b)k + \frac{1}{2!}[f_{x_1x_1}(a,b)h^2 + f_{x_1x_2}(a,b)hk + f_{x_2x_2}(a,b)k^2] + \cdots \tag{12.7}$$

where subscripts denote derivatives, for example, $f_{x_1} = \partial f/\partial x_1$. For the point a, b to be a *stationary* point, the first derivatives with respect to the independent variables must be zero, that is, $f_{x_1} = 0$ and $f_{x_2} = 0$. As was true in the single-variable case, to determine whether this stationary point is a maximum, a minimum, or a *saddle point*, it is necessary to examine the second derivatives. A local maximum occurs when

$$f_{x_1x_1} < 0$$

and

$$f_{x_1x_1}f_{x_2x_2} - (f_{x_1x_2})^2 > 0 \tag{12.8}$$

A local minimum occurs when

$$f_{x_1x_1} > 0$$

and

$$f_{x_1x_1}f_{x_2x_2} - (f_{x_1x_2})^2 > 0 \tag{12.9}$$

For saddle points,

$$f_{x_1x_1}f_{x_2x_2} - (f_{x_1x_2})^2 = 0 \tag{12.10}$$

With reference to Fig. 12.2, location Ⓐ represents a local maximum, as well as the absolute maximum for the range shown. Location Ⓑ is a local maximum that occurs on the boundary. Locations Ⓓ and Ⓔ both represent local minima, and depending on their values, one represents the absolute minimum for the region. Location Ⓒ is a saddle point.

The corresponding necessary conditions for maximization or minimization of a continuous function of n independent variables are

$$\frac{\partial f}{\partial x_1} = 0 \qquad \frac{\partial f}{\partial x_2} = 0 \qquad \cdots \qquad \frac{\partial f}{\partial x_n} = 0 \tag{12.11}$$

By using the method of Leitmann,† a local maximum will be realized if

$$D_n > 0 \qquad n = 2, 4, 6, \ldots \tag{12.12a}$$

or

$$D_n < 0 \qquad n = 1, 3, 5, \ldots \tag{12.12b}$$

where

$$D_n \equiv \begin{vmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n} \end{vmatrix} \tag{12.13}$$

For a minimum

$$D_n > 0 \qquad n = 1, 2, 3, \ldots \tag{12.14}$$

Again, it must be recalled that maximum and minimum values can also occur at locations of discontinuities in the given function(s) or on boundaries.

12.1.1 Constrained Maximum and Minimum Problems

Problems in the general theory of maxima and minima are classified not only according to the number of independent variables involved but also as to whether these variables are interconnected by some other relationship or subsidiary condition. For example, the problem may be to maximize $f(x_1,x_2)$, subject to the constraint $\varphi(x_1,x_2) = 0$. Such a case is illustrated diagrammatically in Fig. 12.3. It should be evident that one direct method for solution of this problem would be to solve the equation $\varphi(x_1,x_2) = 0$ and obtain a relationship between x_1 and x_2. Substitution of this relationship into $f(x_1,x_2)$ will reduce the maximization problem to one of a single independent variable.

In general terms, the constrained maximum (or minimum) problem can be expressed as the maximization of the function

$$f = f(x_1,x_2,x_3, \ldots ,x_n) \tag{12.15}$$

---

† G. Leitmann, The Optimization of Rocket Trajectories: A Survey, in S. F. Singer (ed.), "Progress in the Astronautical Sciences," app. A, North Holland Publishing Company, Amsterdam, 1962.

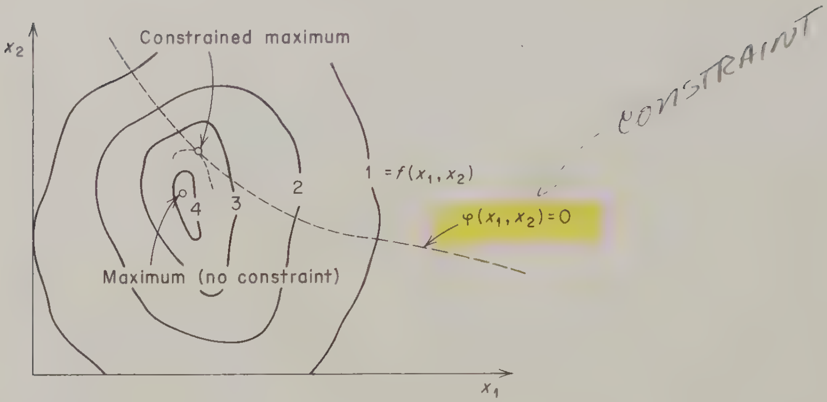


Figure 12.3

subject to

$$\left.\begin{aligned}\varphi_1(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\\ \varphi_2(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\\ \ldots\ldots\ldots\ldots\ldots &\\ \varphi_m(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\end{aligned}\right\} \qquad (12.16)$$

If the m simultaneous equations (12.16) can be solved for m of the independent variables, these can be substituted into Eq. (12.15), thereby reducing the maximization (or minimization) problem to one of $n - m$ independent variables.

$$f = f(x_1,x_2,x_3, \ . \ . \ . \ ,x_{n-m}) \qquad (12.17)$$

Unfortunately, it is often extremely difficult (if not impossible) to obtain explicit expressions for m of the variables contained in Eqs. (12.16). For these types of situations, solution is most often realized by using the method of lagrangian multipliers. This method will be described in detail in Sec. 12.2.

A second general type of constrained maximum (or minimum) problem is found in *control theory*. Here the problem is to maximize the function

$$f = f(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) \qquad (12.18)$$

subject to the m constraining equations

$$\left.\begin{aligned}\varphi_1(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\\ \varphi_2(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\\ \ldots\ldots\ldots\ldots\ldots &\\ \varphi_m(x_1,x_2,x_3, \ . \ . \ . \ ,x_n) &= 0\end{aligned}\right\} \qquad (12.19)$$

and the n equations in p unknown α's

$$\begin{aligned}
\Psi_1(x_1,x_2,x_3, \ldots ,x_n,\alpha_1,\alpha_2,\alpha_3, \ldots ,\alpha_p) &= 0 \\
\Psi_2(x_1,x_2,x_3, \ldots ,x_n,\alpha_1,\alpha_2,\alpha_3, \ldots ,\alpha_p) &= 0 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
\Psi_n(x_1,x_2,x_3, \ldots ,x_n,\alpha_1,\alpha_2,\alpha_3, \ldots ,\alpha_p) &= 0
\end{aligned} \tag{12.20}$$

The problem is to determine the values of α_1, α_2, α_3, . . . , α_p which maximize the value of f.

A third type of constrained maximum (or minimum) problem corresponds to that case where the functional relationships indicated in the second of the cases just described vary with time, and the function to be maximized corresponds to the final values of the independent coordinate variables.

12.1.2 Other Classifications

In addition to classifying optimization problems (1) according to the number of independent variables involved and (2) as to whether or not the optimization is "constrained," it is also necessary to categorize problems as to whether or not they are (3) *static* (i.e., time-independent) or *dynamic* (i.e., time-dependent). The last of the described three constrained maximization problems is time-dependent, whereas the first two are presumed to be static. Still a fourth type of classification is desirable: (4) when no unknown factors are present, the system is referred to as *deterministic*. For a system containing experimental errors and/or other random factors, the problem is *stochastic*.

Before closing this section, it is to be noted that whereas in all the above it has been more or less presumed that the functions to be optimized are algebraic and/or transcendental, many problems require that integral functions be optimized. These quantities depend not upon the values of n discrete variables, but rather upon n complete functions. The problem then is to determine a function (or functions) which maximizes (or minimizes) the integral(s). Again, constraining equations may be specified. These general types of problems are discussed in Sec. 12.3.

12.2 Lagrangian-multiplier Method

The lagrangian-multiplier method is directly applicable and ideally suited to the solution of constrained maximum and minimum problems. Consider the determination of the stationary point(s) for a function

$$\boxed{f = f(x_1,x_2,x_3, \ldots ,x_n)} \tag{12.21}$$

subjected to the subsidiary conditions

$$\text{CONDITIONS}\quad \boxed{\begin{aligned} \varphi_1(x_1,x_2,x_3,\ \ldots\ ,x_n) &= 0 \\ \varphi_2(x_1,x_2,x_3,\ \ldots\ ,x_n) &= 0 \\ \cdots\cdots\cdots\cdots&\cdots \\ \varphi_m(x_1,x_2,x_3,\ \ldots\ ,x_n) &= 0 \end{aligned}} \tag{12.22}$$

To facilitate solution of the problem a new function $\bar{f}$ will first be defined.

$$N = \boxed{\bar{f} = f + \lambda_1\varphi_1 + \lambda_2\varphi_2 + \lambda_3\varphi_3 + \cdots + \lambda_m\varphi_m} \tag{12.23}$$

It is to be noted that the function $\bar{f}$ contains the original function to be maximized, f, each of the constraining equations φ_i, and m undetermined (so-called lagrangian) multipliers λ_i.

For $\bar{f}$ to have a maximum, which will also maximize f, the first partial derivatives with respect to each of the independent variables must equal zero. That is,

$$\boxed{\frac{\partial \bar{f}}{\partial x_1} = 0 \qquad \frac{\partial \bar{f}}{\partial x_2} = 0 \qquad \frac{\partial \bar{f}}{\partial x_3} = 0 \qquad \cdots \qquad \frac{\partial \bar{f}}{\partial x_n} = 0} \tag{12.24}$$

In expanded form, these are

$$\begin{aligned} \frac{\partial f}{\partial x_1} + \lambda_1 \frac{\partial \varphi_1}{\partial x_1} + \lambda_2 \frac{\partial \varphi_2}{\partial x_1} + \lambda_3 \frac{\partial \varphi_3}{\partial x_1} + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1} &= 0 \\ \frac{\partial f}{\partial x_2} + \lambda_1 \frac{\partial \varphi_1}{\partial x_2} + \lambda_2 \frac{\partial \varphi_2}{\partial x_2} + \lambda_3 \frac{\partial \varphi_3}{\partial x_2} + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_2} &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots&\cdots \\ \frac{\partial f}{\partial x_n} + \lambda_1 \frac{\partial \varphi_1}{\partial x_n} + \lambda_2 \frac{\partial \varphi_2}{\partial x_n} + \lambda_3 \frac{\partial \varphi_3}{\partial x_n} + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n} &= 0 \end{aligned} \tag{12.25}$$

These n equations together with the m subsidiary (constraining) equations

$$\begin{aligned} \varphi_1 &= 0 \\ \varphi_2 &= 0 \\ \cdots&\cdots \\ \varphi_m &= 0 \end{aligned} \tag{12.26}$$

allow solution of the $n + m$ unknowns:

$$x_1,\ x_2,\ x_3,\ \ldots\ ,\ x_n,\ \lambda_1,\ \lambda_2,\ \lambda_3,\ \ldots\ ,\ \lambda_m$$

Example 12.1 It is desired to determine the relative cross-sectional areas of the two members in the two-bar pin-connected truss shown in

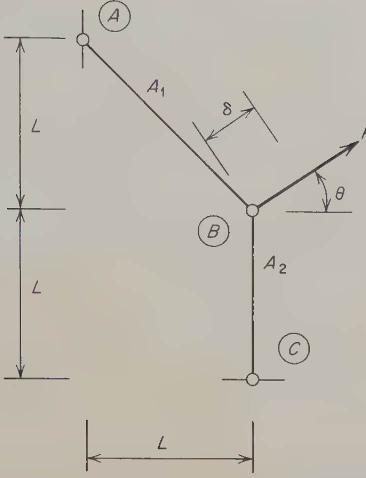


Figure 12.4

Fig. 12.4 which result in the least total weight of the truss. The constraining condition is that the value of the deflection in the direction of the applied load P must equal a preassigned value δ. It is to be presumed that the members are of uniform cross section throughout their lengths, though they may—and probably will—differ one from the other, and that stresses are within the elastic limit of the material.

The cross-sectional area of member AB is A_1. For member BC, the cross-sectional area has been denoted as A_2. Since the weights of each of the members are proportional to their cross-sectional areas times their respective lengths, the objective function to be optimized (minimized) is

$$W(A_1,A_2) = C_1(\sqrt{2}\,A_1 + A_2) \tag{12.27}$$

where W = total weight of truss
$C_1 = \rho L$
ρ = density of material

By using any of the methods of deformable solids, the deflection is found to be

$$\delta = \frac{PL}{E}\left[\frac{2\sqrt{2}\cos^2\theta}{A_1} + \frac{(\cos\theta + \sin\theta)^2}{A_2}\right] \tag{12.28}$$

In the normal form for constraining equations, this would be written as

$$\varphi(A_1,A_2) = \frac{2\sqrt{2}\cos^2\theta}{A_1C_2} + \frac{(\cos\theta + \sin\theta)^2}{A_2C_2} - 1 = 0 \tag{12.29}$$

where
$$C_2 = \frac{\delta E}{PL} \tag{12.30}$$

By proceeding as in Eq. (12.23), the "new (nonconstrained) objective function" to be minimized is

$$\bar{W} = W + \lambda\varphi \tag{12.31}$$

or

$$\bar{W} = C_1(\sqrt{2}\,A_1 + A_2) + \lambda\left[\frac{2\sqrt{2}\cos^2\theta}{A_1C_2} + \frac{(\cos\theta + \sin\theta)^2}{A_2C_2} - 1\right] \tag{12.32}$$

The necessary conditions for optimization are, therefore,

$$\frac{\partial\bar{W}}{\partial A_1} = \sqrt{2}\,C_1 + \lambda\,\frac{-2\sqrt{2}\,(\cos^2\theta)C_2}{(A_1C_2)^2} = 0 \tag{12.33}$$

$$\frac{\partial\bar{W}}{\partial A_2} = C_1 + \lambda\,\frac{-(\cos\theta + \sin\theta)^2C_2}{(A_2C_2)^2} = 0 \tag{12.34}$$

and

$$\varphi = \frac{\partial\bar{W}}{\partial\lambda} = \frac{2\sqrt{2}\cos^2\theta}{A_1C_2} + \frac{(\cos\theta + \sin\theta)^2}{A_2C_2} - 1 = 0 \tag{12.35}$$

Solution of these three simultaneous equations for the necessary design parameters yields

$$A_1C_2 = \sqrt{2}\,(\cos\theta)(3\cos\theta + \sin\theta) \tag{12.36}$$

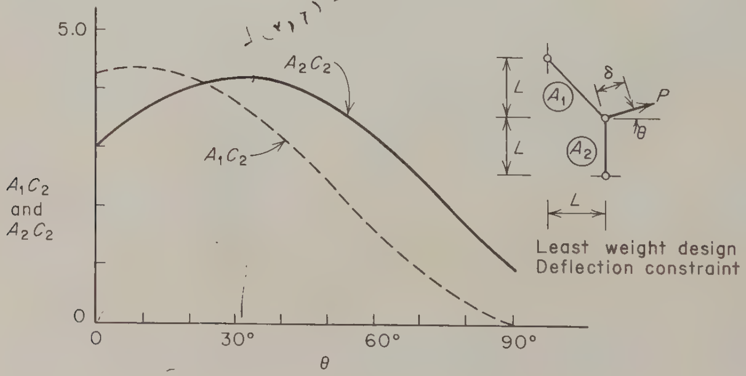
$$A_2C_2 = (\cos\theta + \sin\theta)(3\cos\theta + \sin\theta) \tag{12.37}$$

These are plotted in Fig. 12.5.

Example 12.2 Find the minimum distance from the origin to the curve

$$y = x^2 - 2x + 2 \tag{12.38}$$

Figure 12.5



Here, the problem is to minimize the function

$$f(x,y) = x^2 + y^2 \tag{12.39}$$

subject to the constraining equation

$$\varphi(x,y) \equiv x^2 - 2x + 2 - y = 0 \tag{12.40}$$

If the constraint is incorporated into a new function to be optimized, the augmented equation comparable to (12.23) is

$$\bar{f} = f + \lambda\varphi = x^2 + y^2 + \lambda(x^2 - 2x - y + 2) = 0 \tag{12.41}$$

The necessary conditions for a minimum are, therefore,

$$\frac{\partial \bar{f}}{\partial x} = 2x + \lambda(2x - 2) = 0 \tag{12.42}$$

$$\frac{\partial \bar{f}}{\partial y} = 2y + \lambda(-1) = 0 \tag{12.43}$$

and

$$\frac{\partial \bar{f}}{\partial \lambda} = \varphi = x^2 - 2x - y + 2 = 0 \tag{12.44}$$

From Eqs. (12.42) and (12.43)

$$x = \frac{\lambda}{1 + \lambda} \qquad \text{and} \qquad y = \frac{\lambda}{2} \tag{12.45}$$

Substituting these values into Eq. (12.44) yields

$$\lambda^3 - 3\lambda - 4 = 0 \tag{12.46}$$

The one real root of this cubic equation is

$$\lambda = 2.196 \tag{12.47}$$

which when substituted into Eq. (12.45) yields

$$x = 0.687 \qquad \text{and} \qquad y = 1.098 \tag{12.48}$$

This, in turn, gives the desired minimum distance as

$$\min f(x,y) = \sqrt{x^2 + y^2} = \sqrt{1.678} = 1.295 \tag{12.49}$$

To illustrate that the solution given as Eq. (12.49) is, in fact, the desired minimum, consider the two curves of Fig. 12.6. The first is a plot of the constraining equation. The second, which uses the paired values of x and y from the first, gives the relationship between the coordinate dimension x and the distance r from the origin to the constraining curve.

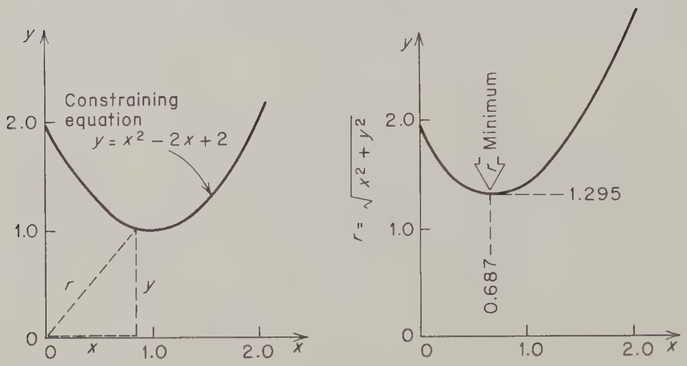


Figure 12.6

12.3 *Introduction to the Calculus of Variation*

The calculus of variation is primarily concerned with the determination of extreme values of functions defined by integrals whose integrands contain one or more functions assuming the role of arguments. That is, the problem deals with the determination of the *extremum* of a function F of functions. Or, in still other words, the calculus of variation is concerned with the establishment of an infinite number of points, all of which identify a curve, a surface, a hypersurface, etc., depending on the nature of the problem, and all of which result in an extreme value for the integral in question.

To illustrate the concept, consider the integral

$$I(y) = \int_a^b F(x,y,y_x,y_{xx})\,dx \tag{12.50}$$

where x is the independent variable, y is an unknown function of x, and $F(x,y,y_x,y_{xx})$ is a known function of the variables x, y, $y_x \equiv dy/dx$, and $y_{xx} \equiv d^2y/dx^2$. (The term *functional* is frequently used to describe functions defined by integrals whose arguments themselves are functions.)

It is to be presumed at the outset that the function $y(x)$ is continuous throughout the interval in question, $a \leq x \leq b$, and is differentiable at least through one more than the order noted in F. It is further assumed that the unknown functions $y(x)$ take on at the boundaries $x = a$ and $x = b$ the known values $y(a)$ and $y(b)$, respectively.

If it is prescribed that $y = y(x)$ is the correct solution which results in an extremum (a maximum or a minimum) for $I(y)$, then it is possible to consider a second function

$$\bar{y}(x) = y(x) + \epsilon\eta(x) \tag{12.51}$$

which results in a greater value for the integral. ϵ is a small, real parameter, whereas $\eta(x)$ is an arbitrary, continuous function defined in the interval $a \le x \le b$ so that the function itself and its first derivative with respect to x vanish on the boundary at $x = a$ and $x = b$. Two possible curves are illustrated in Fig. 12.7. It is to be recognized that in the limit the second functional, $I(y + \epsilon\eta)$, must reduce to the correct solution. That is,

$$\delta I = \frac{dI(y + \epsilon\eta)}{d\epsilon}\bigg|_{\epsilon=0} = 0 \tag{12.52}$$

where δI is the first variation of I. Since

$$I(y + \epsilon\eta) = \int_a^b F(x, y + \epsilon\eta, y_x + \epsilon\eta_x, y_{xx} + \epsilon\eta_{xx})\, dx \tag{12.53}$$

$$\delta I = \frac{dI(y + \epsilon\eta)}{d\epsilon}\bigg|_{\epsilon=0} = \int_a^b (F_y\eta + F_{y_x}\eta_x + F_{y_{xx}}\eta_{xx})\, dx = 0 \tag{12.54}$$

where $$F_y \equiv \frac{\partial F}{\partial y} \qquad F_{y_x} \equiv \frac{\partial F}{\partial y_x} \qquad F_{y_{xx}} \equiv \frac{\partial F}{\partial y_{xx}} \tag{12.55}$$

It is to be noted that as ϵ approaches zero, the first variation of the dependent function $y(x)$ is

$$\delta y = \bar{y} - y = \epsilon\eta \tag{12.56}$$

Similarly,

$$\delta y_x = \epsilon\eta_x$$

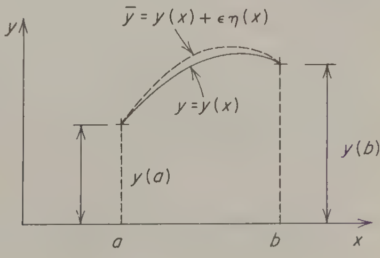
and

$$\delta y_{xx} = \epsilon\eta_{xx} \tag{12.57}$$

Multiplying Eq. (12.54) by ϵ and substituting these values yields

$$\delta I = \int_a^b (F_y\, \delta y + F_{y_x}\, \delta y_x + F_{y_{xx}}\, \delta y_{xx})\, dx = 0 \tag{12.58}$$

Figure 12.7



Integrating, by parts, the last two terms gives

$$\int_a^b (F_{y_x}\,\delta y_x)\,dx = [F_{y_x}\,\delta y]_a^b - \int_a^b \frac{d}{dx}(F_{y_x})\,\delta y\,dx \tag{12.59}$$

$$\int_a^b (F_{y_{xx}}\,\delta y_{xx})\,dx = [F_{y_{xx}}\,\delta y_x]_a^b - \left[\frac{d}{dx}(F_{y_{xx}})\,dy\right]_a^b + \int_a^b \frac{d^2}{dx^2}(F_{y_{xx}})\,\delta y\,dx \tag{12.60}$$

But $\eta(x,y)$ was selected so that $\eta(a) = \eta(b) = 0$ and $\eta_x(a) = \eta_x(b) = 0$. Therefore Eq. (12.58) can be written as

$$\delta I = \int_a^b \left(F_y - \frac{d}{dx}F_{y_x} + \frac{d^2}{dx^2}F_{y_{xx}}\right)\delta y\,dx = 0 \tag{12.61}$$

However, δy can be any small, arbitrary value different from zero. Therefore, the integral can be zero only if the expression contained within the parentheses equals zero.

$$F_y - \frac{d}{dx}F_{y_x} + \frac{d^2}{dx^2}F_{y_{xx}} = 0 \tag{12.62}$$

Equation (12.62) is the necessary *eulerian differential equation* associated with the variational problem

$$I(y) = \min$$

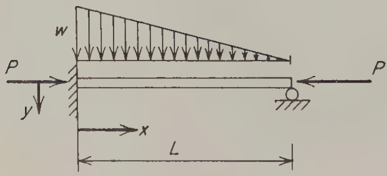
where $$I(y) = \int F(x,y,y_x,y_{xx})\,dx \tag{12.63}$$

That is, to minimize $I(y)$, the unknown function $y(x)$ must satisfy Eq. (12.62).

Example 12.3 The potential energy U of a beam of length L, uniform cross section—having a bending stiffness per unit of length of EI—and subject to the loading shown in Fig. 12.8, is given by

$$U = \int_0^L \left[\frac{EI}{2}(y_{xx})^2 - \frac{P}{2}(y_x)^2 - w\left(1 - \frac{x}{L}\right)y\right]dx \tag{12.64}$$

Figure 12.8



It is required to determine the differential equation corresponding to the equilibrium state.

For the illustrated case the boundary conditions are $y(0) = y_x(0) = 0$ and $y(L) = y_{xx}(L) = 0$.

Since it is known that equilibrium corresponds to that particular $y(x)$ which minimizes the total potential energy of the systems,

$$\delta U = 0 \tag{12.65}$$

will result in the desired solution.

With the notation of Eq. (12.50), the function F is

$$F = \frac{EI}{2}(y_{xx})^2 - \frac{P}{2}(y_x)^2 - w\left(1 - \frac{x}{L}\right)y \tag{12.66}$$

Therefore

$$\begin{aligned} F_y &\equiv \frac{\partial F}{\partial y} = -w\left(1 - \frac{x}{L}\right) \\ F_{y_x} &\equiv \frac{\partial F}{\partial y_x} = -P(y_x) \\ F_{y_{xx}} &\equiv \frac{\partial F}{\partial y_{xx}} = EI(y_{xx}) \end{aligned} \tag{12.67}$$

Still further differentiations yield

$$\frac{d}{dx}(F_{y_x}) = -P(y_{xx})$$

and

$$\frac{d^2}{dx^2}(F_{y_{xx}}) = EI(y_{xxxx}) \tag{12.68}$$

Substituting these values into the Euler equation gives the following fourth-order ordinary differential equation:

$$EI(y_{xxxx}) + P(y_{xx}) = w\left(1 - \frac{x}{L}\right)$$

or

$$\frac{d^4y}{dx^4} + \frac{P}{EI}\frac{d^2y}{dx^2} = \frac{w}{EI}\left(1 - \frac{x}{L}\right) \tag{12.69}$$

This equation can be solved by any of the regular methods of ordinary differential equations, or by the numerical procedures described in Chap. 10. It is to be recognized that the solution must satisfy the given boundary conditions.

The procedures developed above for the solution of functions of one independent variable can be extended to other cases. Consider, for example, the solution to the problem of minimizing the integral

$$I(u) = \iint F(x,y,u,u_x,u_y)\,dx\,dy \tag{12.70}$$

In this case $u \equiv u(x,y)$ and $u_x \equiv \partial u/\partial x$ and $u_y \equiv \partial u/\partial y$.

Again, assuming a correct solution $w(x,y)$ which minimizes the integral in question,

$$I(u + \epsilon\eta) = \iint F(x,y,\ u + \epsilon\eta,\ u_x + \epsilon\eta,\ u_y + \epsilon\eta_y)\,dx\,dy \tag{12.71}$$

or

$$\delta I = \iint_{ac}^{bd} \left(\frac{\partial F}{\partial u}\,\delta u + \frac{\partial F}{\partial u_x}\,\delta u_x + \frac{\partial F}{\partial u_y}\,\delta u_y \right) dx\,dy = 0 \tag{12.72}$$

Possible surfaces are illustrated in Fig. 12.9. If it is required that $\eta(x,y) = 0$ on the boundaries, Eq. (12.72) reduces to

$$\delta I = \iint_{ac}^{bd} \left(\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\frac{\partial F}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial F}{\partial u_y} \right) \delta u\,dx\,dy = 0 \tag{12.73}$$

The eulerian equation corresponding to this case is, therefore,

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) = 0 \tag{12.74}$$

Table 12.1 summarizes for various integrals the eulerian differential equations which yield stationary values of functionals.

Example 12.4 Show that a function that minimizes the integral

$$I = \iint (\nabla^2 u)^2\,dx\,dy \tag{12.75}$$

is the solution of the biharmonic equation

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0 \tag{12.76}$$

Figure 12.9

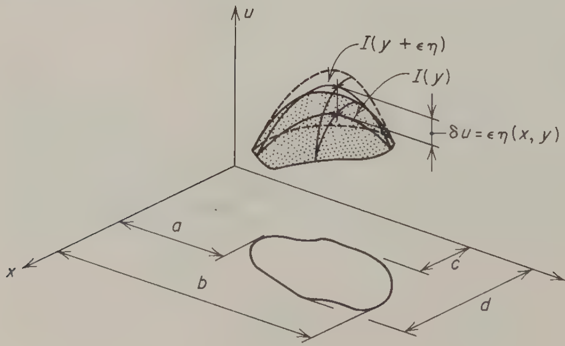


Table 12.1 Eulerian Differential Equations Which Yield Stationary Values for Various Integrals

Integral Equations	Eulerian Differential Equations Corresponding to $\delta I = 0$
One Independent Variable	
$I(y) = \int F(x,y_1,y_2,y_3, \ldots ,y_n)\,dx$	$\frac{\partial F}{\partial y_1} = \frac{\partial F}{\partial y_2} = \frac{\partial F}{\partial y_3} = \cdots = \frac{\partial F}{\partial y_n} = 0$
$I(y) = \int F(x,y,y_x)\,dx$	$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right) = 0$
$I(y) = \int F(x,y,y_x,y_{xx})\,dx$	$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y_{xx}}\right) = 0$
$I(y,z, \ldots) = \int F(x,y,z, \ldots, y_x,z_x, \ldots ,y_{xx},z_{xx}, \ldots)\,dx$	$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_x}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y_{xx}}\right) - \cdots = 0$ $\frac{\partial F}{\partial z} - \frac{d}{dx}\left(\frac{\partial F}{\partial z_x}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial z_{xx}}\right) - \cdots = 0$ . . . .
Two Independent Variables	
$I(u) = \iint F(x,y,u,u_x,u_y)\,dx\,dy$	$\frac{\partial F}{\partial x} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) = 0$
$I(u) = \iint F(x,y,u,u_x,u_y, u_{xx},u_{xy},u_{yy})\,dx\,dy$	$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\partial F}{\partial u_{xx}}\right)$ $+ 2\frac{\partial^2}{\partial x\,\partial y}\left(\frac{\partial F}{\partial u_{xy}}\right) + \frac{\partial^2}{\partial y^2}\left(\frac{\partial F}{\partial u_{yy}}\right) = 0$

Since

$$\nabla^2 u = u_{xx} + u_{yy} \tag{12.77}$$

the function F in Eq. (12.75) is determined by expansion.

$$I = \iint (\nabla^2 u)^2\,dx\,dy = \iint (u_{xx}^2 + 2u_{xx}u_{yy} + u_{yy}^2)\,dx\,dy$$

or

$$F = u_{xx}^2 + 2u_{xx}u_{yy} + u_{yy}^2 \tag{12.78}$$

From Table 12.1, it is noted that the Euler equation required to minimize I (that is, $\delta I = 0$) is

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) + \frac{\partial^2}{\partial x^2}\left(\frac{\partial F}{\partial u_{xx}}\right) + 2\frac{\partial^2}{\partial_x\partial_y}\left(\frac{\partial F}{\partial u_{xy}}\right) + \frac{\partial^2}{\partial y^2}\left(\frac{\partial F}{\partial u_{yy}}\right) = 0 \tag{12.79}$$

But for the functions listed as Eq. (12.78)

$$\frac{\partial F}{\partial u} = 0 \qquad \frac{\partial F}{\partial u_x} = 0 \qquad \frac{\partial F}{\partial u_y} = 0 \qquad \frac{\partial F}{\partial u_{xy}} = 0 \tag{12.80}$$

$$\begin{aligned} \frac{\partial F}{\partial u_{xx}} &= 2u_{xx} + 2u_{yy} = 2(u_{xx} + u_{yy}) \\ \frac{\partial F}{\partial u_{yy}} &= 2u_{xx} + 2u_{yy} = 2(u_{xx} + u_{yy}) \end{aligned} \tag{12.81}$$

Therefore the Euler differential equation yields

$$\frac{\partial^2}{\partial x^2}[2(u_{xx} + u_{yy})] + \frac{\partial^2}{\partial y^2}[2(u_{xx} + u_{yy})] = 0$$

or

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0 \tag{12.82}$$

Constrained extremum problems also exist in many areas of engineering and applied science. The constraints may be either algebraic, differential, or integral equations in form, or they may be a combination of the three.

For the *special case* where n *independent* variables $y_1(x)$, $y_2(x)$, $y_3(x)$, . . . , $y_n(x)$ prescribe a function

$$I(y_1,y_2,y_3, \ldots ,y_n) = \int_a^b F(x,y_1,y_2,y_3, \ldots ,y_n)\,dx \tag{12.83}$$

which is to be optimized, and which is subject to the constraining equations

$$\begin{aligned} J_1 &= \int_a^b \Phi_1(x,y_1,y_2,y_3, \ldots ,y_n) = 0 \\ J_2 &= \int_a^b \Phi_2(x,y_1,y_2,y_3, \ldots ,y_n) = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ J_m &= \int_a^b \Phi_m(x,y_1,y_2,y_3, \ldots ,y_n) = 0 \end{aligned} \tag{12.84}$$

the problem can be solved as a succession of ordinary minimum problems. The procedure will be the same as that described in Sec. 12.3. A new augmented integral $\bar{I}$ first will be defined which contains m undefined, constant lagrangian multipliers.

$$\bar{I} = I + \lambda_1 J_1 + \lambda_2 J_2 + \lambda_3 J_3 + \cdots + \lambda_m J_m \tag{12.85}$$

From this

$$\min \bar{I} = \min \int_a^b (F + \lambda_1\Phi_1 + \lambda_2\Phi_2 + \lambda_3\Phi_3 + \cdots + \lambda_m\Phi_m)\,dx \tag{12.86}$$

But since, by definition, the $y_i(x)$'s are not interrelated, optimization of the integral can be handled as a series of ordinary minimum problems for each value of x from a to b. The necessary conditions are

$$
\begin{aligned}
&\frac{\partial}{\partial y_1}(F + \lambda_1\Phi_1 + \lambda_2\Phi_2 + \lambda_3\Phi_3 + \cdots + \lambda_m\Phi_m) = 0\\
&\frac{\partial}{\partial y_2}(F + \lambda_1\Phi_1 + \lambda_2\Phi_2 + \lambda_3\Phi_3 + \cdots + \lambda_m\Phi_m) = 0\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\frac{\partial}{\partial y_n}(F + \lambda_1\Phi_1 + \lambda_2\Phi_2 + \lambda_3\Phi_3 + \cdots + \lambda_m\Phi_m) = 0
\end{aligned}
\tag{12.87}
$$

For the more general case where the y_i's are not, by definition, independent one from the other but are interrelated through integral constraining equation(s), the following more general procedure must be followed. The functional to be minimized is, for example,

$$I(y_i) = \int_a^b F(x,y_i,\dot{y}_i)\,dx \qquad i = 1, 2, 3, \ldots, n \tag{12.88}$$

The constraining equations are

$$J_j = \int_a^b \Phi_j(x,y_i,\dot{y}_i)\,dx \qquad i = 1, 2, 3, \ldots, n;\ j = 1, 2, 3, \ldots, m \tag{12.89}$$

[In both of these equations the dot sign denotes a derivative with respect to the independent variable; that is, $\dot{y}_i \equiv (d/dx)(y_i)$.]

To facilitate solution there first will be defined the augmented function $\bar{F}$, which is composed of the original F plus the sum of the products of a set of m lagrangian multipliers λ_j and the constraining equation (12.89).

$$\bar{F} = F + \lambda_j\Phi_j \qquad j = 1, 2, 3, \ldots, m \tag{12.90}$$

or

$$\bar{F} = F + \lambda_1\Phi_1 + \lambda_2\Phi_2 + \lambda_3\Phi_3 + \cdots + \lambda_m\Phi_m \tag{12.91}$$

The necessary condition(s) for solution follow directly from application of the appropriate Euler equation

$$\frac{\partial}{\partial y_i}(F + \lambda_j\Phi_j) - \frac{d}{dx}\left[\frac{\partial}{\partial \dot{y}_i}(F + \lambda_j\Phi_j)\right] = 0 \tag{12.92}$$

where $i = 1, 2, 3, \ldots, n$, and $j = 1, 2, 3, \ldots, m$.

When the integral to be optimized is

$$\int_a^b F(x,y_i,\dot{y}_i)\,dx \qquad i = 1, 2, 3, \ldots, n \tag{12.93}$$

and the constraining equations are

$$\varphi_j = \varphi_j(x,y_i) = 0 \qquad i = 1, 2, 3, \ldots, n;\ j = 1, 2, 3, \ldots, m \tag{12.94}$$

the augmented function $\bar{F}$ is defined as in Eq. (12.95).

$$\bar{F} = F + \mu_j \varphi_j \qquad j = 1, 2, 3, \ldots, m \tag{12.95}$$

Here, μ_j is a lagrangian-type multiplier, which, instead of being a constant, is a function of the independent variable. That is,

$$\mu_j = \mu_j(x) \tag{12.96}$$

The necessary conditions for solution are therefore the appropriate Euler equations,

$$\frac{\partial \bar{F}}{\partial y_i} - \frac{d}{dx}\left(\frac{\partial \bar{F}}{\partial \dot{y}_i}\right) = 0 \qquad i = 1, 2, 3, \ldots, n \tag{12.97}$$

and the constraining equations (12.94).

For the still more general set of equations

$$I(y,z) = \int_a^b F(x,y,z,\dot{y},\dot{z})\, dx \tag{12.98}$$

with the constraints

$$\varphi(x,y,z,\dot{y},\dot{z}) = 0 \tag{12.99}$$

the Euler-type equations required to ensure $\delta I = 0$ are

$$\begin{aligned} &\frac{\partial}{\partial y}(F) - \frac{d}{dx}\left(\frac{\partial F}{\partial \dot{y}}\right) - \left[\mu \frac{\partial \varphi}{\partial y} - \frac{d}{dx}\left(\mu \frac{\partial \varphi}{\partial \dot{y}}\right)\right] = 0 \\ \text{and}\quad &\frac{\partial}{\partial z}(F) - \frac{d}{dx}\left(\frac{\partial F}{\partial \dot{z}}\right) - \left[\mu \frac{\partial \varphi}{\partial z} - \frac{d}{dx}\left(\mu \frac{\partial \varphi}{\partial \dot{z}}\right)\right] = 0 \end{aligned} \tag{12.100}$$

These in combination with Eq. (12.99) provide a sufficient number of conditions to define y, z, and μ as functions of x.

Example 12.5 A flexible rope of uniform cross section and length L hangs in equilibrium under the influence of gravity. Its two ends are attached to two fixed points A and B, as indicated in Fig. 12.10. The problem is to determine the equation of the curve assumed by the rope.

If g is the acceleration of gravity and ρ is the mass per unit length along the rope, the potential energy of the entire rope with respect to the datum $y = 0$ is

$$I(y) = \int_0^L (\rho g) y\, ds \tag{12.101}$$

The constraint required is that the integral of the ds elements equal the total length of the rope. That is,

$$L = \int_0^L ds \tag{12.102}$$

By noting that

$$ds = \sqrt{(dx)^2 + (dy)^2} = dx \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \tag{12.103}$$

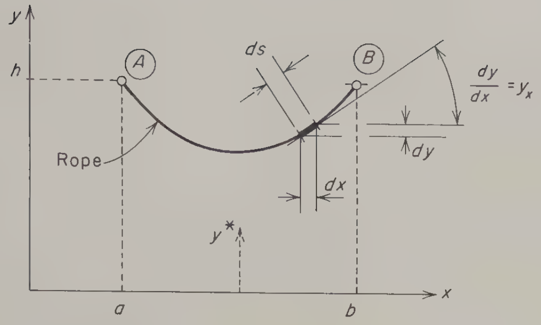


Figure 12.10

Eqs. (12.101) and (12.102) can be written as

$$I(y) = \rho g \int_a^b y(1 + y_x^2)^{1/2}\, dx \tag{12.104}$$

and

$$L = \int_a^b (1 + y_x^2)^{1/2}\, dx \tag{12.105}$$

To ensure minimization, the Euler differential equation to be solved is

$$(1 + y_x^2)^{1/2} - \frac{d}{dx}\left[\frac{(y - \lambda)y_x}{(1 + y_x^2)^{1/2}}\right] = 0 \tag{12.106}$$

Integration yields the desired equation of a catenary:

$$y = \lambda + (\zeta) \cosh \frac{x + \xi}{\zeta} \tag{12.107}$$

To facilitate evaluation of the constants λ, ζ, and ξ, a new location of the y coordinate axis, one which bisects the distance ba, will first be presumed. (This is shown as y^* in Fig. 12.10.) Since it is specified that points A and B have the same $y = h$ values,

$$h - \lambda = (\zeta) \cosh \frac{-(a + b)/2 + \xi}{\zeta} = (\zeta) \cosh \frac{+(a + b)/2 + \xi}{\zeta} \tag{12.108}$$

or $\xi = 0$. Solution of Eq. (12.105), using the value for y listed in Eq. (12.107), yields

$$L = 2(\zeta) \sinh \frac{(a + b)/2}{\zeta} \tag{12.109}$$

From this, with L and $(a + b)/2$ known, ζ can be determined.

Example 12.6 Considering the family of curves that lie wholly on a given surface S and pass through points A and B, determine the differential equations required to minimize the arc length from A to B. (See Fig. 12.11.)

The total distance on the surface between the two given points A and B is

$$I = \int_a^b \sqrt{x_t^2 + y_t^2 + z_t^2}\, dt \tag{12.110}$$

where each of the coordinate dimensions is related to a fourth (independent) parameter t. That is, $x = x(t)$, $y = y(t)$, and $z = z(t)$. Since the arc must lie on the surface S, the constraining equation is

$$S(x,y,z) = 0 \tag{12.111}$$

The presumed boundary conditions correspond to $t = a$ and $t = b$.

Proceeding as described in Eq. (12.95) gives the augmented function to be optimized as

$$\bar{F} = F + \mu S \tag{12.112}$$

where

$$\begin{aligned} F &= (x_t^2 + y_t^2 + z_t^2)^{1/2} \\ \mu &= \mu(t) \\ S &= S(x,y,z) \end{aligned} \tag{12.113}$$

Equations (12.97) yield the following three equations:

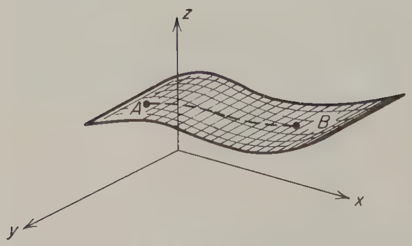
$$\mu \frac{\partial S}{\partial x} - \frac{d}{dt}\left(\frac{x_t}{F}\right) = 0 \qquad \mu \frac{\partial S}{\partial y} - \frac{d}{dt}\left(\frac{y_t}{F}\right) = 0$$

$$\mu \frac{\partial S}{\partial z} - \frac{d}{dt}\left(\frac{z_t}{F}\right) = 0 \tag{12.114}$$

Eliminating μ from each of these gives the desired two equations

$$\frac{\dfrac{d}{dt}\left(\dfrac{x_t}{F}\right)}{\dfrac{\partial S}{\partial x}} = \frac{\dfrac{d}{dt}\left(\dfrac{y_t}{F}\right)}{\dfrac{\partial S}{\partial y}} = \frac{\dfrac{d}{dt}\left(\dfrac{z_t}{F}\right)}{\dfrac{\partial S}{\partial z}} \tag{12.115}$$

Figure 12.11



These, together with the particular equation of the surface (12.111), allow solution of the problem in question.

If the surface is a sphere whose center is at the origin, the constraining equation is

$$S(x,y,z) = x^2 + y^2 + z^2 - a^2 = 0 \tag{12.116}$$

and Eqs. (12.115) reduce to

$$\frac{Fx_{tt} - x_tF_t}{2xF^2} = \frac{Fy_{tt} - y_tF_t}{2yF^2} = \frac{Fz_{tt} - z_tF_t}{2zF^2} \tag{12.117}$$

By rearranging terms, these can be written as

$$\frac{yx_{tt} - xy_{tt}}{yx_t - xy_t} = \frac{F_t}{F} = \frac{zy_{tt} - yz_{tt}}{zy_t - yz_t} \tag{12.118}$$

Considering only the first and last group of terms in this equation and integrating twice yields

$$x - py + qz = 0 \tag{12.119}$$

which is the equation of a plane through the center of the sphere. That is, the minimum arc lies on a *great circle* through points A and B.

12.3.1 Ritz's Method

While boundary-value and other types of problems can be described in terms of integral expressions which have minimum values for those specific functions which solve the particular problems in question, many do not admit of any rigorous solution. For those which do, it has been demonstrated that solution is equivalent to obtaining answers to appropriate Euler differential equations. However, an equally important use of the principles that have been developed relates to the construction of sequences of approximate functions which converge to the desired Euler solutions and therefore to the resolution of an appropriate variational problem. Ritz's method is one such method that presumes that the solution can be approximated by a linear combination of suitably chosen functions.

To facilitate description of Ritz's method, it will be presupposed that the exact solution for a particular extremum problem corresponds to $y(x) = y^*(x)$. That is, $I(y^*)$ corresponds to, for example, a minimum value of I. If a new "admissible function" $\bar{y}(x)$ is substituted into the functional equation,

$$I(\bar{y}) \geq I(y^*) \tag{12.120}$$

Moreover, if a convergent sequence of functions

$$\bar{y}_n(x) = \bar{y}_1 + \bar{y}_2 + \cdot \cdot \cdot + \bar{y}_n$$

is selected such that $\bar{y}_n(x)$ contains the complete family

$$\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_{n-1}$$

it is reasonable to conclude that

$$\lim_{n\to\infty} \bar{y}_n = y^* \tag{12.121}$$

Ritz proposed the construction of a linear combination of functions of the general type

$$\bar{y}(x) = \varphi(x,a_1,a_2, \ldots ,a_k) \tag{12.122}$$

where φ is such that for all values of a_i the end conditions are satisfied. $a_1, a_2, \ldots, a_k$ are real parameters.

Substitution of Eq. (12.122) into the functional equation to be optimized yields

$$\bar{I}(a_1,a_2,a_3, \ldots ,a_k) = 0 \tag{12.123}$$

which can be handled directly, as described in the earlier parts of this chapter. That is,

$$\frac{\partial \bar{I}}{\partial a_j} = 0 \qquad j = 1, 2, \ldots, k \tag{12.124}$$

results in minimization (or maximization) of the now algebraic function $\bar{I}$. It is to be noted that the sequence of families of functions is

$$\begin{aligned} y_1(x) &= \varphi_1(x,a_1) \\ y_2(x) &= \varphi_2(x,a_1,a_2) \\ y_3(x) &= \varphi_3(x,a_1,a_2,a_3) \\ &\cdots\cdots\cdots\cdots\cdots \\ y_n(x) &= \varphi_n(x,a_1,a_2,a_3, \ldots ,a_n) \end{aligned} \tag{12.125}$$

Convergence to the correct solution depends primarily on the selection of "relatively complete sets of functions" for the φ's. Two relatively complete sets in the interval $0 \le x \le \ell$, which satisfy boundary conditions $y(0) = y(\ell) = 0$, are

$$\bar{y}_k(x) = \sum_{k=1}^{n} a_k x^k(\ell - x) \tag{12.126}$$

and

$$\bar{y}_k(x) = \sum_{k=1}^{n} a_k \sin \frac{k\pi x}{\ell} \tag{12.127}$$

It is to be noted that the Ritz method requires that only the "geometrical boundary conditions" (y and y_x) be satisfied by the approximating function.

Example 12.7 By using Ritz's method, the extremum of the functional

$$I(y) = \int_0^1 [(y_x)^2 - xy]\, dx \tag{12.128}$$

subject to the boundary conditions $y(0) = 0$ and $y(1.0) = 0$, is to be determined.

Assume as an approximating polynomial the following power series:

$$\bar{y}_k(x) = \sum_{k=1}^{n} (1 - x)(a_k x^k) \tag{12.129}$$

It is to be observed that both boundary conditions are simultaneously satisfied by this choice of functions. Table 12.2 summarizes several of the terms required in the solution.

Table 12.2

n	$\bar{y}_n$	$\bar{y}_{n_x}$
1	$(1 - x)(a_1x)$	$a_1(1 - 2x)$
2	$(1 - x)(a_1x + a_2x^2)$	$a_1(1 - 2x) + a_2(2x - 3x^2)$
3	$(1 - x)(a_1x + a_2x^2 + a_3x^3)$	$a_1(1 - 2x) + a_2(2x - 3x^2) + a_3(3x^2 - 4x^3)$

Consider the first possible case, $n = 1$: Substitution of the appropriate terms into Eq. (12.128) yields

$$I(\bar{y}_1) = \int_0^1 \{[a_1(1 - 2x)]^2 - x[a_1(x - x^2)]\}\, dx \tag{12.130}$$

Integrating and substituting in the limits give

$$I(\bar{y}_1) = \tfrac{1}{3}a_1^2 - \tfrac{1}{12}a_1 \tag{12.131}$$

Since this equation is an algebraic expression in the variable a_1, it is possible to handle the minimization in an elementary fashion

$$\delta I = 0 = \frac{\partial I}{\partial a_1} = \tfrac{2}{3}a_1 - \tfrac{1}{12} = 0 \tag{12.132}$$

or

$$a_1 = \tfrac{1}{8} \tag{12.133}$$

The corresponding function for $\bar{y}_1(x)$ is then

$$\bar{y}_1(x) = (1 - x)(\tfrac{1}{8}x) \tag{12.134}$$

For the case where $n = 2$, the functional is

$$I(\bar{y}) = \int_0^1 \{[a_1(1 - 2x) + a_2(2x - 3x^2)]^2 - x[a_1(x - x^2) + a_2(x^2 - x^3)]\}\,dx \qquad (12.135)$$

Integrating and substituting in the limits yield

$$I(\bar{y}) = \tfrac{1}{3}a_1^2 + \tfrac{1}{3}a_1a_2 + \tfrac{2}{15}a_2^2 - \tfrac{1}{12}a_1 - \tfrac{1}{20}a_2 \qquad (12.136)$$

For $\delta I = 0$,

$$\frac{\partial I(\bar{y})}{\partial a_1} = \tfrac{2}{3}a_1 + \tfrac{1}{3}a_2 - \tfrac{1}{12} = 0$$

and

$$\frac{\partial I(\bar{y})}{\partial a_2} = \tfrac{1}{3}a_1 + \tfrac{4}{15}a_2 - \tfrac{1}{20} = 0 \qquad (12.137)$$

Solving these simultaneously gives

$$a_1 = \tfrac{1}{12} \qquad a_2 = \tfrac{1}{12} \qquad (12.138)$$

The equation for the approximating function is therefore

$$\bar{y}_2(x) = (1 - x)(\tfrac{1}{12}x + \tfrac{1}{12}x^2) = \tfrac{1}{12}(1 - x)(x - x^2) \qquad (12.139)$$

Additional terms can be taken into account; however, it is to be recognized that the number of required computations increases at an increasing rate.

12.3.2 Galerkin's Method

In general, Galerkin's method employs a so-called "semi-inverse" procedure for solution of extremum problems. A differential equation is first presumed (or known) which corresponds to a Euler equation for an appropriate functional. For example, if the differential equation selected is

$$L(y) = p(x)[y] - [q(x)] - \frac{d}{dx}[r(x)y_x] = 0 \qquad (12.140)$$

where $p(x)$, $q(x)$, and $r(x)$ are presumed to be functions of x or constants, it can be demonstrated that the functional being optimized is

$$I(y) = \int F(x,y,y_x)\,dx \qquad (12.141)$$

where

$$F = \tfrac{1}{2}r(x)y_x^2 + \tfrac{1}{2}p(x)y^2 - q(x)y \qquad (12.142)$$

In Example 12.4 it was shown that solution of

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0 \qquad (12.143)$$

is equivalent to maximizing (or minimizing) the functional

$$I(y) = \iint (\nabla^2 u)^2 \, dx \, dy \tag{12.144}$$

For the more general nonhomogeneous fourth-order differential equation

$$\nabla^4 u = f(x,y) \tag{12.145}$$

the functional is

$$I(y) = \iint [(\nabla^2 u)^2 - fu] \, dx \, dy \tag{12.146}$$

It is to be recalled that the basic concept of the Galerkin method requires that the error associated with an assumed polynomial be orthogonal to a set of linearly independent functions φ_i. (See Sec. 10.4.) That is,

$$\int_X \epsilon \varphi_i \, dx = 0 \tag{12.147}$$

where ϵ = error due to substitution of trial polynomial into given differential equation

φ_i = suitable coordinate functions (which satisfy *all* the boundary conditions associated with the differential equation in question)

The trial polynomial is presumed to be of the general form

$$\bar{y} = \sum_{i=1}^{n} a_i \varphi_i \tag{12.148}$$

where the a_i's are constants.

Substitution of the finite polynomial described in Eq. (12.148) into the given differential equation $L(y) = 0$ will ordinarily result in a value other than zero; that is, there will be an error.

$$L(\bar{y}) = \epsilon(x) \qquad \epsilon \not\equiv 0 \tag{12.149}$$

The general Galerkin method [Eq. (12.147)] therefore requires

$$\int_X L\left(\sum_{j=1}^{n} a_j \varphi_j\right) \varphi_i \, dx = 0 \qquad i = 1, 2, \ldots, n \tag{12.150}$$

The subscript change within the parentheses from i to j is to emphasize that the error is a function of the total polynomial assumed, whereas orthogonalization takes place independently with respect to each of the coordinate functions.

Example 12.8 Using Galerkin's method, again determine the extremum of the functional

$$I(y) = \int_0^1 [(y_x)^2 - xy] \, dx \tag{12.151}$$

subject to the boundary conditions $y(0) = y(1.0) = 0$. The associated Euler equation is

$$L(y) = y_{xx} + \tfrac{1}{2}x = 0 \tag{12.152}$$

The selected polynomial approximation (which satisfies the given boundary conditions, term by term) is

$$\bar{y}_k = \sum_{k=1}^{n} (1 - x)(a_k x^k) \tag{12.153}$$

By considering just the first term in this series,

$$\bar{y}_1 = a_1 x(1 - x) \tag{12.154}$$

the error is

$$L(\bar{y}) = \bar{y}_{xx} + \tfrac{1}{2}x = -2a_1 + \tfrac{1}{2}x = \epsilon \tag{12.155}$$

The coordinate function is determined as follows:

$$\bar{y}_1 = a_1\varphi_1 = a_1[x(1 - x)]$$

or

$$\varphi_1 = x(1 - x) \tag{12.156}$$

The necessary Galerkin equation is

$$\int_0^1 (-2a_1 + \tfrac{1}{2}x)[x(1 - x)]\,dx = 0 \tag{12.157}$$

which yields

$$a_1 = \tfrac{1}{8} \tag{12.158}$$

For the case where two terms are presumed,

$$\bar{y}_2 = a_1 x(1 - x) + a_2 x^2(1 - x) \tag{12.159}$$

the error is

$$\epsilon = a_1(-2) + a_2(2 - 6x) + \tfrac{1}{2}x \tag{12.160}$$

and the coordinate functions are

$$\varphi_1 = x(1 - x) \tag{12.161}$$

and

$$\varphi_2 = x^2(1 - x) \tag{12.162}$$

Here, two Galerkin equations are required:

$$\int_0^1 [a_1(-2) + a_2(2 - 6x) + \tfrac{1}{2}x][x(1 - x)]\,dx = 0 \tag{12.163}$$

$$\int_0^1 [a_1(-2) + a_2(2 - 6x) + \tfrac{1}{2}x][x^2(1 - x)]\,dx = 0 \tag{12.164}$$

From the first of these,

$$8a_1 + 4a_2 = 1 \tag{12.165}$$

From the second,

$$20a_1 + 16a_2 = 3 \tag{12.166}$$

Simultaneous solution yields

$$a_1 = \tfrac{1}{12} \qquad a_2 = \tfrac{1}{12} \tag{12.167}$$

These are exactly the same values that were obtained by Ritz's method. This is not a coincidence. For many (if not most) variational problems in engineering, Galerkin's method (with the assumed coordinate functions satisfying *all* the prescribed boundary conditions) and Ritz's method (with only "essential" boundary conditions being met by the coordinate functions) result in the same solution.[†] Moreover, it is to be recognized that Galerkin's rules for obtaining the coefficients of the coordinate functions are simpler than Ritz's rules.

12.4 Direct vs. Indirect Methods

By definition, an indirect method for optimization is one that is based on a "necessary condition" for that optimization. A direct method, on the other hand, depends on a "direct" comparison of function values at two (or more) points. To illustrate the basic difference in these two approaches, consider the problem of determining the maximum value of the single-independent-variable problem

$$y = f(x) = \text{maximum} \tag{12.168}$$

An indirect approach would be, for example, to determine the value of x for which $y_x = 0$. The direct approach is concerned with establishing, by comparison of various y values, the particular value of x for which y is a maximum. All the methods previously described in this chapter are indirect methods.

When analytical solutions are desired, indirect methods are preferred. For individual cases, however, direct methods are more often used. Direct methods have a further advantage: end points and points of discontinuity in the function value present no additional computational difficulties. It is to be recognized also that as the number of independent variables increases, the direct method has a further computational advantage: the indirect method requires the evaluation of at least n equations, whereas the direct method requires the successive evaluation of only one equation.

Direct methods for optimization are most often referred to as *seek methods* or *search methods*. For single-variable problems, where it is known that the function is unimodal, these procedures have been extensively developed, and it is possible to prescribe patterns of search that will result in the definition of known "intervals of uncertainty" within a minimum number of trials. A few of the more often referred to single-variable search techniques are the *Fibonacci search, the golden section, lattice search*, etc. Multidimensional search problems are not so clearly understood. Nevertheless, usable procedures such as *steepest ascent*

[†] L. Collatz, "The Numerical Treatment of Differential Equations," p. 208, Springer-Verlag OHG, Berlin, 1960. (Translated from a supplemental version of the second German edition by P. G. Williams.)

(or *descent*), *sectioning*, *pattern search*, and others have been developed to handle these types of situations. Section 12.5 is concerned specifically with single-variable search problems. Section 12.6 deals with the *direct method* of solution of multidimensional problems.

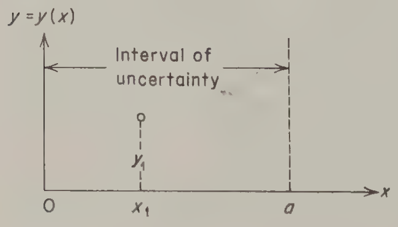
12.5 Single-variable Search

Throughout this section it is presumed that the particular function being examined is *unimodal* in the interval in question. [See Eqs. (12.4).] Moreover, it is assumed that sequential search procedures—those procedures where information from one trial can be used to benefit the location of subsequent trials—are to be developed, rather than simultaneous ones—where all trials are run at once.

To define the general problem and illustrate the procedure that will be used in its solution, consider a range of independent variable $0 \leq x \leq a$. Nothing is presumed known in advance concerning the function in question except that it is unimodal. Assume now that the function value at a particular value $x = x_1$ is determined (for example, by evaluating a given function or by carrying out an experiment). This is shown in Fig. 12.12. With only this much information known, it is impossible to determine whether the maximum (or minimum) value of y occurs to the left, to the right, or at x_1. Therefore, the "interval of uncertainty" in the location of that particular x^* for which y is a maximum is the entire interval from $x = 0$ to $x = a$.

If now a second trial is carried out at, say, $x = x_2 > x_1$, and if the resulting value $y_2 = y(x_2) > y_1 = y(x_1)$, the interval of uncertainty can be reduced from that previously established. The definition of unimodality [Eqs. (12.4)] rules out the possibility that $y_{\max}$ can exist to the left of x_1. Therefore, $y_{\max}$ must exist within the interval $x_1 \leq x^* \leq a$, as shown in Fig. 12.13. Should a third trial $x_2 \leq x_3 \leq a$ result in a value of y_3 less than y_2, it can be similarly reasoned that $y_{\max}$ must be contained within the interval $x_1 \leq x^* \leq x_3$. This is shown in Fig. 12.14.

Figure 12.12



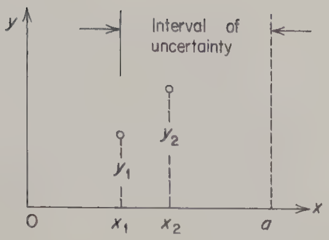


Figure 12.13

It is the purpose of this section to develop procedures for progressively and methodically reducing the interval of uncertainty, and thereby allow the determination of x^* and $y_{\max}$ with a minimum number of trials.

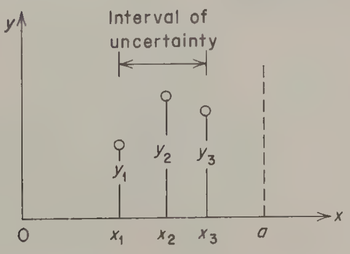
12.5.1 The Search Plan

For the problem just described, depending on the values of the y's, it was possible to select new trial values of x which allow a progressive "zeroing-in" on the desired optimum value. It is to be recognized, however, that no real, analytically defined, repeating pattern of search was followed.

A more meaningful search plan would be one that allows an a priori determination of the effectiveness of the search: a plan that is not dependent on the values of the particular dependent function being optimized, one that is concerned only with defining the range of x within which the optimum will occur.

In developing such a search plan, it must be recognized that at any stage in the process the optimum value of the dependent variable will be contained within a "one-division x interval" on either side of the then known maximum trial value of y. For example, in Fig. 12.13 the maximum trial value y_2 is located at x_2. The true maximum $y_{\max}$ therefore

Figure 12.14



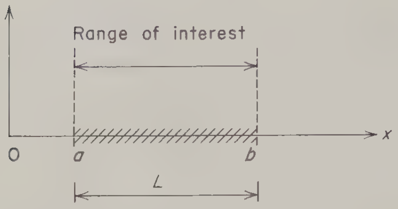
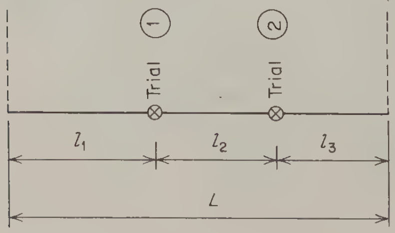


Figure 12.15

must be contained within the interval $x_1 \le x^* \le x_2$, to the left of x_2, or $x_2 \le x^* \le a$, to the right of x_2. Thus the total interval of uncertainty is $x_1 \le x^* \le a$. However, if the values y_1 and y_2 are not known, and if only locations x_1 and x_2 are established, the interval of uncertainty will be either $0 \le x^* \le x_2$ (presuming $y_1 > y_2$) or $x_1 \le x^* \le a$ (presuming $y_2 > y_1$). Since it can be reasoned that in the comparison of various search plans, or the examination of the convergence of a particular plan, it is not the location of the interval but rather the magnitude of the interval that is important, a conservative assumption, suitable for evaluating the effectiveness of a particular search plan, is to select the greatest of the possible intervals of uncertainty. (This is not to say that when a particular search is carried out the wrong segment should be retained for further subdivision. The procedure of defining the worst possible case is selected only to allow a conservative, quantitative evaluation to be made of the search plan itself, one that is independent of the particular variations in the dependent function.)

In Fig. 12.12 it was demonstrated that the selection of only one trial value, regardless of where that trial might be located, will not allow a reduction in the interval of uncertainty. Two trials, however, will effect a reduction. The problem, then, is the determination of where these two trials should be placed. It is presumed that the range of variable of interest is $a \le x \le b$, where $b - a = L$ (see Fig. 12.15). The segment in question is shown enlarged in Fig. 12.16. For the case

Figure 12.16



shown, two possible intervals of uncertainty exist:

1. If $y_1 > y_2$, the interval is $\{L_1\} = \ell_1 + \ell_2$. (12.169)
2. If $y_1 < y_2$, the interval is $\{L_2\} = \ell_2 + \ell_3$. (12.170)

The constraining equations for this elementary minimization problem,

$$\{L_1\} = \ell_1 + \ell_2 = \text{minimum} \tag{12.171}$$

$$\{L_2\} = \ell_2 + \ell_3 = \text{minimum} \tag{12.172}$$

are

$$\ell_1 + \ell_2 + \ell_3 = L \tag{12.173}$$

and

$$\ell_1 > 0 \qquad \ell_2 > 0 \qquad \ell_3 > 0 \tag{12.174}$$

Solving Eq. (12.173) for ℓ_2 and substituting the resulting value into Eqs. (12.171) and (12.172) yields

$$L - \ell_3 = \text{minimum} \tag{12.175}$$

and

$$L - \ell_1 = \text{minimum} \tag{12.176}$$

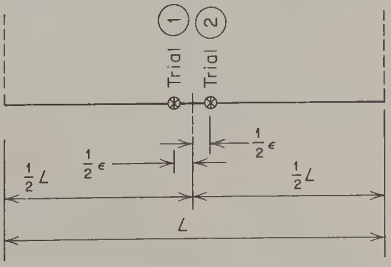
A direct formal solution of these equations suggests that

$$\begin{aligned} \ell_1 &= \ell_3 = \tfrac{1}{2}L \\ \text{with} \qquad \ell_2 &= 0 \end{aligned} \tag{12.177}$$

But the equation $\ell_2 = 0$ does not satisfy the second of the constraining equations (12.174). That is, the solution defined by Eqs. (12.177) requires that both trials be run at the same value of x. This, in turn, requires that the interval of uncertainty be the full value $\{L\}$ and not $\{L_1\}$ or $\{L_2\}$ as presumed in Eqs. (12.169) and (12.170). The correct placement of the two trials for $\{L_1\} = \{L_2\} = \text{minimum}$ is shown in Fig. 12.17, where ϵ is any small positive value.

Since a paired set of experiments has been carried out as indicated in Fig. 12.17, and it has been established that the optimum in question exists to the right (or to the left) of the centerline of the total region, a

Figure 12.17



second set of paired trials can now be run about the three-quarter point (or one-quarter point). Paired trials about an eighth point would next be performed, and so forth. An indication of the effectiveness of this *paired-trial* method, or *dichotomous search,* as it is frequently called, is given in Table 12.3.

Table 12.3

Number of Trials	Interval of Uncertainty
0	L
2	$0.500L$
4	$0.250L$
6	$0.125L$
8	$0.0625L$
10	$0.0313L$
12	$0.0156L$

12.5.2 The Golden Section

In the paired-trial search procedure just described, optimization was required at each and every stage. Such an assumption, however, does not necessarily guarantee the greatest search efficiency. A nonoptimal placement of "early" trial values may well result in greater efficiencies at later stages. A search pattern based on *the golden section* (or the golden ratio†) is one such method that will provide a greater search efficiency than the dichotomous search. The principle of the golden section can be stated as follows:

> *In the division of a line into two unequal parts, the ratio of the larger of the two segments to the total length of the line should be the same as the ratio of the smaller to the larger segment.*

Consider the case shown in Fig. 12.18. The total length of the line in question is ℓ. The unequal segments are a and b, where $a > b$.

From the definition of the golden section,

$$\frac{\ell}{a} = \frac{a}{b} = \frac{1}{\tau} \tag{12.178}$$

where τ is a constant. Since

$$a + b = \ell \tag{12.179}$$

---

† The golden section or golden ratio was known to Greek mathematician Euclid, fl.ca. 300 B.C. It has had a profound influence on art and architecture, because of its connection with esthetics and visual beauty.

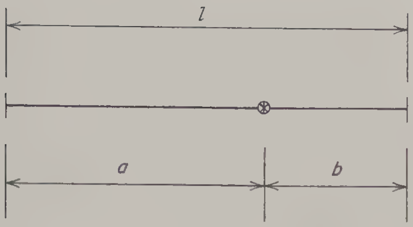


Figure 12.18

substitution of the relationships given in Eq. (12.178) yields

$$\tau\ell + \tau a = \tau\ell + \tau(\tau\ell) = \ell$$

or

$$\tau^2 + \tau - 1 = 0 \qquad (12.180)$$

The single positive root of this equation is

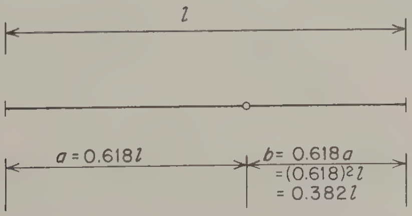
$$\tau = +0.618 \qquad (12.181)$$

The proportions of the golden section therefore are those shown in Fig. 12.19.

To illustrate the procedure to be followed in a golden-section search, consider the sequence illustrated in Fig. 12.20. The initial interval is presumed to be of unit length, as shown in Fig. 12.20*a*. The first trial would be placed as in Fig. 12.20*b*, at a distance of 0.618ℓ from the end. This single experiment, however, does not reduce the interval of uncertainty, as has been indicated earlier in this section.

Since it is impossible at this stage to know whether the optimum in question lies to the left or to the right of the trial $x = 0.618$, the second trial (see Fig. 12.20*c*) should be so placed that the intervals of uncertainty are the same, $\{0.618\}$. If it is assumed that $y(0.618) > y(0.382)$, the interval for further study is contained within the limits $0.382 \leq x^* \leq 1.000$. As shown in Fig. 12.20*d*, to afford equal intervals of uncertainty, the third experiment would be symmetrically placed at $x = 0.764$. In a similar fashion the fourth trial would be at $x = 0.854$ (presuming the optimum lies to the right-hand side of the segment in question). For

Figure 12.19



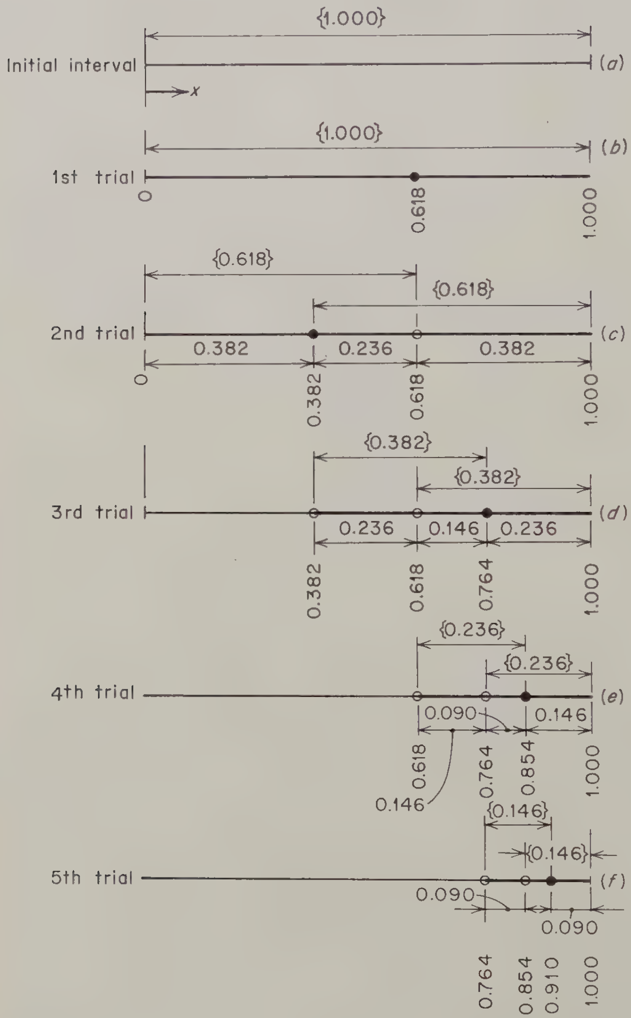


Figure 12.20

the fifth trial, $x = 0.910$. For the sixth, $x = 0.944$. Table 12.4 summarizes the intervals of uncertainty corresponding to the number of golden-section trials. These should be compared with those listed in Table 12.3 for the paired-trial search.

12.5.3 Fibonacci Search

It is to be noted that while the golden-section search described above is demonstrably more efficient than the dichotomous (paired-trial) plan for

Table 12.4

Number of Trials	Interval of Uncertainty
0	L
1	L
2	$0.618L$
3	$0.382L$
4	$0.236L$
5	$0.146L$
6	$0.090L$
7	$0.056L$
8	$0.0345L$
9	$0.0213L$
10	$0.0132L$
11	$0.0081L$
12	$0.0050L$

$n > 2$, no optimizing, as such, was employed to obtain a "best" possible interval of uncertainty. The Fibonacci scheme with which this discussion is concerned assumes a desired optimum after n trials.

In developing this search plan it must first be recognized that symmetrical placement of trials within a given segment results in greatest reduction in subsequent intervals of uncertainty. (Note, for example, Fig. 12.20.) Moreover, optimum placement of the last two experiments requires that they be an ϵ distance one from the other, symmetrically placed about the centerline of the next-to-the-last interval. A search pattern can therefore be deduced by reasoning backward from these last experiments, as illustrated in Fig. 12.21. In all cases the trial point in question is shown as a solid circle. The interval of uncertainty is designated by the brackets; for example, $\{L_n\}$ is the interval of uncertainty after the nth trial. Again, it should be understood that while for illustrative purposes it has been presumed that the optimum occurs near the right-hand end of the original segment, the same reasoning and results would follow from any other assumption.

Comparing the lengths of the intervals shown in Fig. 12.21*a*, *b*, and *c* reveals a relationship between any three consecutive intervals of uncertainty. For the case illustrated

$$\{L_{n-2}\} = \{L_{n-1}\} + \{L_n\} \tag{12.182}$$

For the more general case where $1 < i < n$

$$\{L_{i-2}\} = \{L_{i-1}\} + \{L_i\} \tag{12.183}$$

Again, working backward from the last trial (note the geometry of Fig.

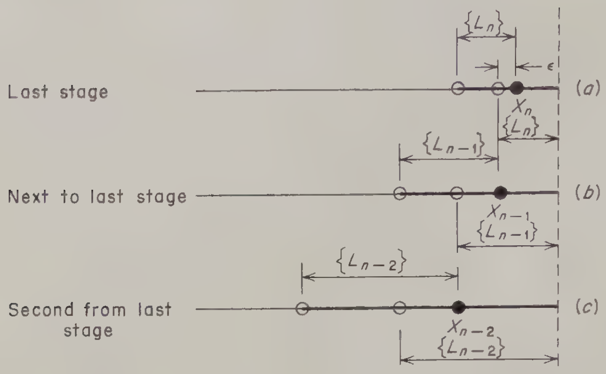


Figure 12.21

12.21*a*),

$$\{L_{n-1}\} = 2\{L_n\} - \epsilon \tag{12.184}$$

Substitution of this into Eq. (12.182) yields

$$\{L_{n-2}\} = [2\{L_n\} - \epsilon] + \{L_n\} = 3\{L_n\} - \epsilon \tag{12.185}$$

The interval $\{L_{n-3}\}$ would be defined as follows:

$$\begin{aligned}\{L_{n-3}\} &= \{L_{n-2}\} + \{L_{n-1}\} \\ &= [3\{L_n\} - \epsilon] + [2\{L_n\} - \epsilon] \\ &= 5\{L_n\} - 2\epsilon\end{aligned} \tag{12.186}$$

In similar fashion

$$\{L_{n-4}\} = 8\{L_n\} - 3\epsilon$$

and

$$\{L_{n-5}\} = 13\{L_n\} - 5\epsilon \tag{12.187}$$

and so forth. These equations can be written in more compact form in terms of so-called Fibonacci numbers, where

$$\{L_{n-k}\} = F_{k+1}\{L_n\} - F_{k-1}\epsilon \tag{12.188}$$

and

$$\begin{aligned}F_0 &= 1 \\ F_1 &= 1 \\ F_k &= F_{k-1} + F_{k-2} \qquad k = 2, 3, \ldots\end{aligned} \tag{12.189}$$

The first 12 Fibonacci numbers are listed in Table 12.5.

If it is assumed that the initial length to be investigated is 1.00, the interval of uncertainty after n trials can be determined from Eq. (12.188).

$$\{L_n\} = \frac{1}{F_n} + \frac{F_{n-2}}{F_n}(\epsilon) \tag{12.190}$$

Table 12.5

k	F_k
0	1
1	1
2	2
3	3
4	5
5	8
6	13
7	21
8	34
9	55
10	89
11	144
12	233

The first experiment should be located at

$$x_1 = \frac{F_{n-1}}{F_n} + (-1)^n \frac{\epsilon}{F_n} \tag{12.191}$$

Table 12.6 lists the interval of uncertainty associated with the Fibonacci method as a function of the number of trials to be carried out.

Table 12.6

Number of Trials	Interval of Uncertainty
0	L
1	L
2	$0.500L$
3	$0.333L$
4	$0.200L$
5	$0.125L$
6	$0.077L$
7	$0.048L$
8	$0.0294L$
9	$0.0182L$
10	$0.0112L$
11	$0.0069L$
12	$0.0043L$

12.6 Multivariable Search

All the single-variable search methods developed in Sec. 12.5 presume a unimodal function. While such an assumption may be a reasonable one for many single-variable engineering-type problems, it is much less likely to be the case when the function in question is dependent on more than one variable. The value of a direct extension of the methods and reasoning just described is therefore questionable.

Of even greater concern is the fact that optimization along a line—any line—in an n-dimensional space does not ensure that the value so obtained is an optimum for the total space, even though the function in question behaves in a unimodal manner along the line being examined. This fact suggests that for multidimensional problems, instead of behaviors along lines being considered, regions within the total space be examined. A direct comparison of the optimum function values within each region could then be used to determine that a particular region be excluded from further consideration, or that it exclude the comparison region. Such methods have been developed. However, they will not be discussed here. Instead, a more direct continuous procedure will be developed.

12.6.1 The One-at-a-time, or Sectioning, Method

A logical extension of the one-dimensional case discussed in Sec. 12.5 would be to select a starting point—any starting point—for the search and, holding all but one of the independent variables constant, alter that one until an optimum is realized. Then, in a manner not unlike that of the Gauss-Seidel iterative method for solving sets of algebraic equations, holding constant the just-determined independent variable and all others but one at their previously assigned values, allow a second variable to seek its relative optimal value. In turn, by altering only one variable at a time, holding all others constant, all independent quantities are allowed to vary. The procedure is continued until no future improvement is realized. This is illustrated for a two-independent-variable problem in Fig. 12.22, where $f = f(x_1,x_2)$. In the figure, contours of equal values of f are shown. The particular search plan followed is shown by arrows.

The method just described is known as *sectional search* or the *one-at-a-time method*. When the function in question is convex, and when it is symmetrical about a pair of axes parallel to the coordinate axes, it is an extremely efficient search technique. This is illustrated in Fig. 12.23*a*. The case where the axes are at angles to the coordinate directions is shown in Fig. 12.23*b*. Here, the method results in the correct solutions being found, but more steps are involved.

For certain types of functions, and certain starting locations for the

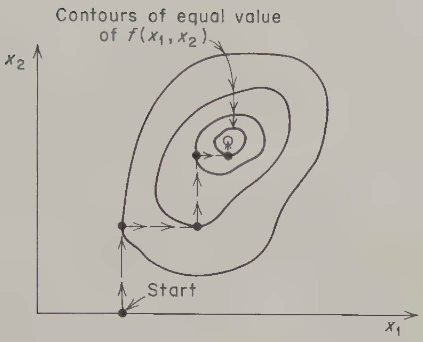


Figure 12.22

search, the one-at-a-time method can yield completely false locations for the optimum solution. One such case is illustrated in Fig. 12.24. Therefore, unless more information concerning the general behavior of the function is known in advance, sectional search is not recommended.

12.6.2 The Method of Steepest Descent (or Ascent)

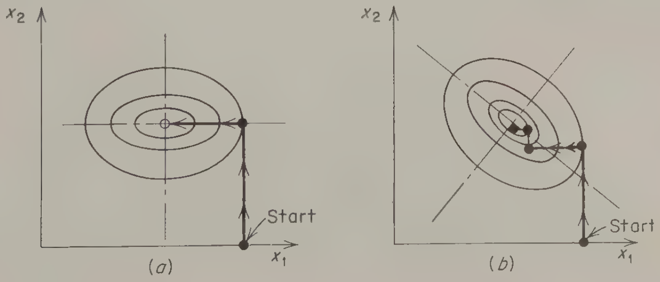
Also known as the *method of gradients*, the *method of steepest descent* has found wide acceptance as a powerful multidimensional seek procedure. It can be used for both nonconstrained and constrained types of problems, as well as for the solution of extremum situations.

Consider first the problem of defining in an open domain the optimum of a function

$$f = f(x_1,x_2, \ . \ . \ . \ ,x_n) \tag{12.192}$$

It is presumed that the function possesses continuous partial derivatives with respect to each of the independent variables, and that the minimum value of f is desired.

Figure 12.23



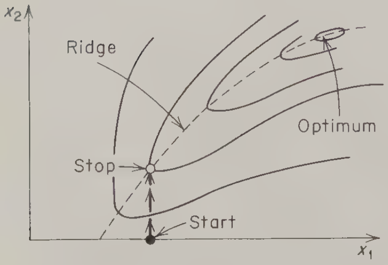


Figure 12.24

From a starting point $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$, a ds change is presumed, where

$$(ds)^2 = (dx_1)^2 + (dx_2)^2 + \cdots + (dx_n)^2$$

or

$$1 - \sum_{i=1}^{n} \left(\frac{dx_i}{ds}\right)^2 = 0 \tag{12.193}$$

This change, however, alters the value of the function f.

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \cdots + \frac{\partial f}{\partial x_n} dx_n \tag{12.194}$$

With respect to the total coordinate relocation ds, the rate of change in the function value is

$$\frac{df}{ds} = \frac{\partial f}{\partial x_1}\frac{dx_1}{ds} + \frac{\partial f}{\partial x_2}\frac{dx_2}{ds} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{ds}$$

or

$$\frac{df}{ds} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{dx_i}{ds} \tag{12.195}$$

The (dx_i/ds)'s are direction cosines associated with the variation.

In formal terms, the problem at issue is the minimization of a function [Eq. (12.195)], subject to a constraining equation (12.193). This can be accomplished by using the lagrangian-multiplier method discussed in Sec. 12.2. The augmented function is

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{dx_i}{ds} + \lambda\left[1 - \sum_{i=1}^{n}\left(\frac{dx_i}{ds}\right)^2\right] \tag{12.196}$$

where λ is the constant lagrangian multiplier. (It is to be recognized that $\partial f/\partial x_i$ in Eq. (12.196) can be presumed known.) The unknowns are the n particular direction cosines dx_i/ds, which result in the greatest rate of reduction (for steepest descent) in the function value, and the lagrangian multiplier.

Differentiating each of Eqs. (12.196) with respect to dx_i/ds and equating the resulting expression to zero yields

$$\frac{\partial f}{\partial x_i} + \lambda\left(-2\frac{dx_i}{ds}\right) = 0 \qquad i = 1, 2, \ldots, n \tag{12.197}$$

These, together with the constraining equation (12.193), allow solution of the unknown direction cosines and the lagrangian multiplier.

$$\begin{aligned} \frac{dx_1}{ds} &= \pm\frac{\partial f}{\partial x_1}\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{-\frac{1}{2}} \\ \frac{dx_2}{ds} &= \pm\frac{\partial f}{\partial x_2}\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{-\frac{1}{2}} \\ &\cdots\cdots\cdots\cdots \\ \frac{dx_n}{ds} &= \pm\frac{\partial f}{\partial x_n}\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{-\frac{1}{2}} \end{aligned} \tag{12.198}$$

and

$$\lambda = \pm\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{\frac{1}{2}} \tag{12.199}$$

With these equations, the rate of change in the function value can be determined by substitution in Eq. (12.195).

$$\frac{df}{ds} = \pm\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{\frac{1}{2}} \tag{12.200}$$

(In all these equations, the plus values following the equals signs correspond to steepest ascent, whereas negative values are for steepest descent.)

Consider now the introduction of a time parameter t, and *presume that motion along the direction of steepest descent is a continuous process of time.* Since

$$\frac{ds}{dt} = \left[\sum_{i=1}^{n}\left(\frac{dx_i}{dt}\right)^2\right]^{\frac{1}{2}} \tag{12.201}$$

and

$$\frac{dx_i}{dt} = \frac{dx_i}{ds}\frac{ds}{dt} \tag{12.202}$$

$$\frac{dx_j}{dt} = -\frac{\partial f}{\partial x_j}\frac{ds}{dt}\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{-\frac{1}{2}} \tag{12.203}$$

If it is now presumed that the magnitude of ds/dt is proportional to df/ds, the expressions become relatively simple.

$$\frac{ds}{dt} = K\left[\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2\right]^{1/2} \tag{12.204}$$

and

$$\begin{aligned} \frac{dx_1}{dt} &= -K\left(\frac{\partial f}{\partial x_1}\right) \\ \frac{dx_2}{dt} &= -K\left(\frac{\partial f}{\partial x_2}\right) \\ &\cdots\cdots\cdots \\ \frac{dx_n}{dt} &= -K\left(\frac{\partial f}{\partial x_n}\right) \end{aligned} \tag{12.205}$$

The point $(x_1,x_2, \ldots ,x_n)$ moves according to this system of equations.

As an alternative to the continuous procedure just described, a stepwise approach, one that corrects a set of approximations to the solution $\partial f/\partial x_i = 0$ by increments proportional to the negative of $\partial f/\partial x_i$, can be defined. From Eqs. (12.205),

$$\frac{dx_i}{dt} = -K\left(\frac{\partial f}{\partial x_i}\right) \qquad i = 1, 2, \ldots, n \tag{12.206}$$

For the discontinuous process, this would correspond to

$$\Delta x_i = -K\left(\frac{\partial f}{\partial x_i}\right)\Delta t \tag{12.207}$$

By noting ${}_px_i$ as the value of the coordinate x_i corresponding to the pth step in the iterative process, Eq. (12.207) can be written in the more usable form

$${}_{p+1}x_i = {}_px_i - {}_p\left(\frac{\partial f}{\partial x_i}\right)(K\,\Delta t) \qquad i = 1, 2, \ldots, n \tag{12.208}$$

Example 12.9 Determine the minimum value of the function

$$f(x_1,x_2) = 3x_1^2 + x_2^2 \tag{12.209}$$

presuming the starting value $x_1 = 12$, $x_2 = 6$. Use both the continuous and the stepwise approach.

CONTINUOUS SOLUTION

From Eqs. (12.205), presuming $K = 1$,

$$\frac{dx_1}{dt} = -\frac{\partial f}{\partial x_1} = -6x_1$$

and

$$\frac{dx_2}{dt} = -\frac{\partial f}{\partial x_2} = -2x_2 \tag{12.210}$$

These, plus the given starting value (12,6), yield the solutions

$$x_1 = 12e^{-6t} \qquad x_2 = 6e^{-2t} \tag{12.211}$$

Since as $t \to \infty$, both x_1 and $x_2 \to 0$, where $\partial f/\partial x_1 = \partial f/\partial x_2 = 0$, the minimum value of the function corresponds to

$$x_1 = x_2 = f(x_1,x_2) = 0 \tag{12.212}$$

STEPWISE APPROACH

For illustration, it will be presumed that $K = 1.0$ and $\Delta t = \frac{1}{10}$. The solution for the first 10 trials is shown in Table 12.7. The iterative function is

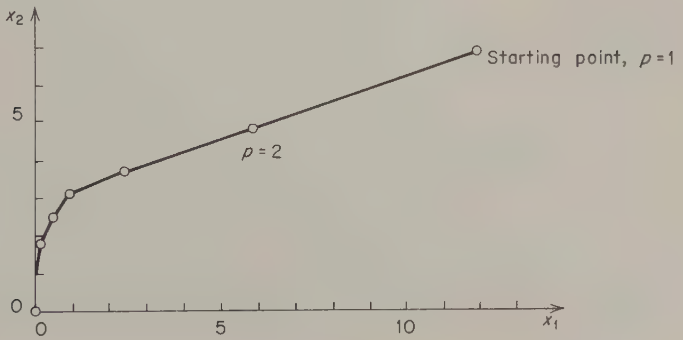
$$_{p+1}x_i = {}_px_i - \frac{1}{10}\,{}_p\!\left(\frac{\partial f}{\partial x_i}\right) \qquad i = 1, 2 \tag{12.213}$$

Table 12.7

P	$_px_1$	$_p\left(\frac{\partial f}{\partial x_1}\right)$	$_{p+1}x_1$	$_px_2$	$_p\left(\frac{\partial f}{\partial x_2}\right)$	$_{p+1}x_2$
1	12	72	5.8	6	12	4.8
2	5.8	34.8	2.32	4.8	9.6	3.84
3	2.32	13.92	0.928	3.84	7.68	3.072
4	0.928	5.568	0.371	3.072	6.144	2.458
5	0.371	2.226	0.148	2.458	4.916	1.966
6	0.148	0.888	0.059	1.966	3.932	1.573
7	0.059	0.354	0.024	1.573	3.146	1.258
8	0.024	0.144	0.010	1.258	2.516	1.006
9	0.010	0.060	0.004	1.006	2.012	0.805
10	0.004	0.024	0.002	0.805	1.610	0.644

The path of convergence is shown in Fig. 12.25.

Figure 12.25



Consider next that particular case where a function

$$f = f(x_1,x_2, \ . \ . \ . \ ,x_n) \tag{12.214}$$

is to be optimized, subject to the given set of constraining equations

$$\varphi_j(x_1,x_2, \ . \ . \ . \ ,x_n) \qquad j = 1, 2, \ . \ . \ . \ , m \tag{12.215}$$

It is to be recognized that the additional geometrical constraint [see Eq. (12.193)]

$$1 - \sum_{i=1}^{n} \left(\frac{dx_i}{ds}\right)^2 = 0 \tag{12.216}$$

must also be included. To facilitate inclusion in the optimization, Eqs. (12.215) are normally expressed in terms of the rate of change with respect to the distance ds. That is,

$$\frac{d\varphi_j}{ds} = \sum_{i=1}^{n} \frac{\partial\varphi_i}{\partial x_i}\frac{dx_i}{ds} = 0 \qquad j = 1, 2, \ . \ . \ . \ , m \tag{12.217}$$

The augmented function to be optimized contains $m + 1$ lagrangian multipliers.

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{dx_i}{ds} + \lambda_0\left[1 - \sum_{i=1}^{n}\left(\frac{dx_i}{ds}\right)^2\right] + \sum_{j=i}^{m} \lambda_j \sum_{i=1}^{n} \frac{\partial\varphi_j}{\partial x_i}\frac{dx_i}{ds} \tag{12.218}$$

Differentiating, partially, each of these equations with respect to (dx_i/ds) and equating the resulting expression to zero yield

$$\frac{\partial f}{\partial x_i} - 2\lambda_0\left(\frac{dx_i}{ds}\right) + \sum_{j=1}^{m} \lambda_j\left(\frac{\partial\varphi_j}{\partial x_i}\right) = 0 \qquad i = 1, 2, \ . \ . \ . \ , n \tag{12.219}$$

These, together with the $m + 1$ constraining equations (12.216) and (12.217), allow solution for the n direction cosines, λ_0 and the m multipliers $\lambda_1, \lambda_2, \ . \ . \ . \ , \lambda_m$.

$$\lambda_0 = \pm\frac{1}{2}\left\{\sum_{i=1}^{m}\left[\frac{\partial f}{\partial x_i} + \sum_{j=1}^{m}\lambda_j\left(\frac{\partial\varphi_j}{\partial x_i}\right)\right]^2\right\}^{1/2} \tag{12.220}$$

$$\frac{dx_i}{ds} = \pm\frac{(\partial f/\partial x_i) + \sum_{j=1}^{m}\lambda_j(\partial\varphi_j/\partial x_i)}{\left\{\sum_{i=1}^{n}\left[(\partial f/\partial x_i) + \sum_{j=1}^{m}\lambda_j(\partial\varphi_j/\partial x_i)\right]^2\right\}^{1/2}} \tag{12.221}$$

Substitution of Eqs. (12.221) into Eqs. (12.217) allows determination of $\lambda_1, \lambda_2, \ . \ . \ . \ , \lambda_m$. The $m \times m$ set of equations to be solved is shown in Table 12.8.

Again, as in the preceding case, a continuous time parameter t will be presumed to control "motion" along the direction of steepest descent. [See Eqs. (12.201) to (12.203).] By taking ds/dt proportional to the denominator of Eq. (12.221), the direction cosines are obtained as

$$\begin{aligned}
\frac{dx_1}{dt} &= -K\left[\frac{\partial f}{\partial x_1} + \lambda_1\left(\frac{\partial \varphi_1}{\partial x_1}\right) + \lambda_2\left(\frac{\partial \varphi_2}{\partial x_1}\right) + \cdots + \lambda_m\left(\frac{\partial \varphi_m}{\partial x_1}\right)\right] \\
\frac{dx_2}{dt} &= -K\left[\frac{\partial f}{\partial x_2} + \lambda_1\left(\frac{\partial \varphi_1}{\partial x_2}\right) + \lambda_2\left(\frac{\partial \varphi_2}{\partial x_2}\right) + \cdots + \lambda_m\left(\frac{\partial \varphi_m}{\partial x_2}\right)\right] \qquad (12.222) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\frac{dx_n}{dt} &= -K\left[\frac{\partial f}{\partial x_n} + \lambda_1\left(\frac{\partial \varphi_1}{\partial x_n}\right) + \lambda_2\left(\frac{\partial \varphi_2}{\partial x_n}\right) + \cdots + \lambda_m\left(\frac{\partial \varphi_m}{\partial x_n}\right)\right]
\end{aligned}$$

It is to be noted that the λ's indicated in (12.222) are to be obtained from solution of the equation given in Table 12.8.

A stepwise approach for steepest descent would require an iterative solution of Eq. (12.223).

$$_{p+1}x_i = {}_px_i - {}_p\left[\frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \lambda_j\left(\frac{\partial \varphi_j}{\partial x_i}\right)\right](K\,\Delta t) \qquad i = 1, 2, \ldots, n \qquad (12.223)$$

Here again, the λ_j's must be obtained from the equations in Table 12.8.

Example 12.10 For the structure and loading shown in Fig. 12.26, determine the cross-sectional areas corresponding to minimum weight of total truss, when the deflection in the direction of the applied load is constrained to equal δ. (See Example 12.1.)

With the notation of Example 12.1, the objective function (weight) to be optimized is

$$W = C_1(\sqrt{2}\,A_1 + A_2)$$

or

$$f(x_1,x_2) = \sqrt{2}\,x_1 + x_2 \qquad (12.224)$$

where

$$x_1 = A_1C_2 \qquad \text{and} \qquad x_2 = A_2C_2 \qquad (12.225)$$

The deflection constraint equation is

$$\varphi(A_1,A_2) = \frac{2\sqrt{2}}{A_1C_2} + \frac{1}{A_2C_2} - 1 = 0 \qquad (12.226a)$$

or

$$\varphi(x_1,x_2) = \frac{2\sqrt{2}}{x_1} + \frac{1}{x_2} - 1 = 0 \qquad (12.226b)$$

Table 12.8

For one constraining equation: $\varphi_1(x_1,x_2,\ .\ .\ .\ ,x_n) = 0$

$$\left(\frac{\partial\varphi_1}{\partial x_1}\frac{\partial\varphi_1}{\partial x_1} + \frac{\partial\varphi_1}{\partial x_2}\frac{\partial\varphi_1}{\partial x_2} + \cdots + \frac{\partial\varphi_1}{\partial x_n}\frac{\partial\varphi_1}{\partial x_n}\right)[\lambda_1] = \left(-\frac{\partial\varphi_1}{\partial x_1}\frac{\partial f}{\partial x_1} - \frac{\partial\varphi_1}{\partial x_2}\frac{\partial f}{\partial x_2} - \cdots - \frac{\partial\varphi_1}{\partial x_n}\frac{\partial f}{\partial x_n}\right)$$

For two constraining equations: $\varphi_1(x_1,x_2,\ .\ .\ .\ ,x_n) = 0$

$\varphi_2(x_1,x_2,\ .\ .\ .\ ,x_n) = 0$

$$\begin{bmatrix} \left(\frac{\partial\varphi_1}{\partial x_1}\frac{\partial\varphi_1}{\partial x_1} + \frac{\partial\varphi_1}{\partial x_2}\frac{\partial\varphi_1}{\partial x_2} + \cdots + \frac{\partial\varphi_1}{\partial x_n}\frac{\partial\varphi_1}{\partial x_n}\right) & \left(\frac{\partial\varphi_1}{\partial x_1}\frac{\partial\varphi_2}{\partial x_1} + \frac{\partial\varphi_1}{\partial x_2}\frac{\partial\varphi_2}{\partial x_2} + \cdots + \frac{\partial\varphi_1}{\partial x_n}\frac{\partial\varphi_2}{\partial x_n}\right) \\ \left(\frac{\partial\varphi_2}{\partial x_1}\frac{\partial\varphi_1}{\partial x_1} + \frac{\partial\varphi_2}{\partial x_2}\frac{\partial\varphi_1}{\partial x_2} + \cdots + \frac{\partial\varphi_2}{\partial x_n}\frac{\partial\varphi_1}{\partial x_n}\right) & \left(\frac{\partial\varphi_2}{\partial x_1}\frac{\partial\varphi_2}{\partial x_1} + \frac{\partial\varphi_2}{\partial x_2}\frac{\partial\varphi_2}{\partial x_2} + \cdots + \frac{\partial\varphi_2}{\partial x_n}\frac{\partial\varphi_2}{\partial x_n}\right) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \left(-\frac{\partial\varphi_1}{\partial x_1}\frac{\partial f}{\partial x_1} - \frac{\partial\varphi_1}{\partial x_2}\frac{\partial f}{\partial x_2} - \cdots - \frac{\partial\varphi_1}{\partial x_n}\frac{\partial f}{\partial x_n}\right) \\ \left(-\frac{\partial\varphi_2}{\partial x_1}\frac{\partial f}{\partial x_1} - \frac{\partial\varphi_2}{\partial x_2}\frac{\partial f}{\partial x_2} - \cdots - \frac{\partial\varphi_2}{\partial x_n}\frac{\partial f}{\partial x_n}\right) \end{bmatrix}$$

Similar matrix equations could be written for larger numbers of constraining equations. The mth equation of these would be

$$\left(\frac{\partial\varphi_m}{\partial x_1}\frac{\partial\varphi_1}{\partial x_1} + \frac{\partial\varphi_m}{\partial x_2}\frac{\partial\varphi_1}{\partial x_2} + \cdots + \frac{\partial\varphi_m}{\partial x_n}\frac{\partial\varphi_1}{\partial x_n}\right)[\lambda_1] + \left(\frac{\partial\varphi_m}{\partial x_1}\frac{\partial\varphi_2}{\partial x_1} + \frac{\partial\varphi_m}{\partial x_2}\frac{\partial\varphi_2}{\partial x_2} + \cdots + \frac{\partial\varphi_m}{\partial x_n}\frac{\partial\varphi_2}{\partial x_n}\right)[\lambda_2]$$

$$+ \cdots + \left(\frac{\partial\varphi_m}{\partial x_1}\frac{\partial\varphi_m}{\partial x_1} + \frac{\partial\varphi_m}{\partial x_2}\frac{\partial\varphi_m}{\partial x_2} + \cdots + \frac{\partial\varphi_m}{\partial x_n}\frac{\partial\varphi_m}{\partial x_n}\right)[\lambda_m] = \left(-\frac{\partial\varphi_m}{\partial x_1}\frac{\partial f}{\partial x_1} - \frac{\partial\varphi_m}{\partial x_2}\frac{\partial f}{\partial x_2} - \cdots - \frac{\partial\varphi_m}{\partial x_n}\frac{\partial f}{\partial x_n}\right)$$

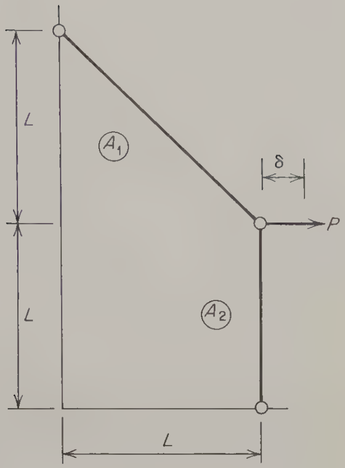


Figure 12.26

The various partial derivatives required in the solution are

$$\frac{\partial \varphi}{\partial x_1} = -\frac{2\sqrt{2}}{x_1^2} \qquad \frac{\partial \varphi}{\partial x_2} = -\frac{1}{x_2^2}$$
$$\frac{\partial f}{\partial x_1} = \sqrt{2} \qquad \frac{\partial f}{\partial x_2} = 1 \tag{12.227}$$

Since there is only one specified constraining equation, there is only one lagrangian multiplier to be determined. Using the first equation of Table 12.8,

$$\left[\left(-\frac{2\sqrt{2}}{x_1^2}\right)\left(-\frac{2\sqrt{2}}{x_1^2}\right) + \left(-\frac{1}{x_2^2}\right)\left(-\frac{1}{x_2^2}\right)\right]\lambda_1 = -\left(-\frac{2\sqrt{2}}{x_1^2}\right)(\sqrt{2}) - \left(-\frac{1}{x_2^2}\right)(1)$$

or

$$\left(\frac{8}{x_1^4} + \frac{1}{x_2^4}\right)\lambda_1 = \frac{4}{x_1^2} + \frac{1}{x_2^2}$$

or

$$\lambda_1 = \frac{4 + \alpha^2}{8 + \alpha^4} x_1^2 \tag{12.228}$$

where

$$\alpha = \frac{x_1}{x_2} \tag{12.229}$$

By presuming a stepwise approach, the equations for iteration are

$$
\begin{aligned}
{}_{p+1}x_1 &= {}_px_1 - {}_p\!\left[\sqrt{2} + \lambda_1\left(-\frac{2\sqrt{2}}{x_1^2}\right)\right](K\,\Delta t) \\
{}_{p+1}x_2 &= {}_px_2 - {}_p\!\left[1 + \lambda_1\left(-\frac{1}{x_2^2}\right)\right](K\,\Delta t)
\end{aligned}
\qquad (12.230)
$$

Substituting in the value of λ_1 from Eq. (12.228),

$$
\begin{aligned}
{}_{p+1}x_1 &= {}_px_1 - \sqrt{2}\;{}_p\!\left(1 - 2\,\frac{4+\alpha^2}{8+\alpha^4}\right)(K\,\Delta t) \\
{}_{p+1}x_2 &= {}_px_2 - {}_p\!\left(1 - \alpha^2\,\frac{4+\alpha^2}{8+\alpha^4}\right)(K\,\Delta t)
\end{aligned}
\qquad (12.231)
$$

Starting values for the seek routine must satisfy the given constraining equation (12.226*b*). For example the following values are chosen:

$$
\begin{aligned}
{}_1x_1 &= 4\sqrt{2} \\
{}_1x_2 &= 2
\end{aligned}
\qquad (12.232)
$$

Table 12.9 summarizes the first five steps of the solution. It is to be noted that $(K\,\Delta t)$ has been arbitrarily selected equal to 1.

Table 12.9

P	${}_px_1$	${}_px_2$	$\alpha = \dfrac{x_1}{x_2}$	$K\,\Delta t$	${}_{p+1}x_1$	${}_{p+1}x_2$
0					5.656	2.000
1	5.656	2.000	2.828	1.000	4.713	2.333
2	4.712	2.333	2.020	↓	4.225	2.671
3	4.240	2.673	1.586		4.112	2.817
4	4.113	2.815	1.461		4.081	2.858
5	4.079	2.860	1.426		4.071	2.871

One of the major difficulties in using the method of steepest descent is the selection of the magnitude of $K\,\Delta t$. The smaller the value, the more closely the stepwise approximation will approach the continuous solution. As the number of constraints increases, however, the determination of the lagrangian multipliers, and thereby the gradient direction, becomes more involved—at an increasing rate. Therefore the exploitation of each step has been suggested as a desirable approach to use in the solution of these types of problems. That is, at each trial point ${}_px_i$, the direction of steepest descent is determined by using the values at that point. *ds* variation in that steepest-descent direction is then continuously increased until a rate of change of zero is noted in the objective function. At that new location, a new gradient direction is determined, and the entire process is repeated. Such a procedure is similar to the one-at-a-time method described earlier, with the exception that the directions are now gradient directions, and are not necessarily parallel to the coordinates.

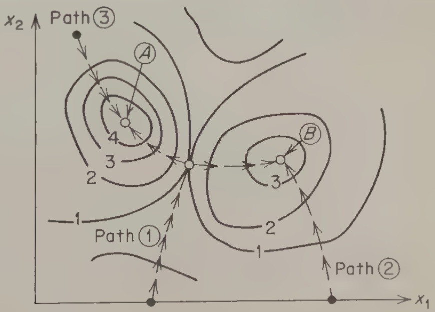
Gradient methods have an inherent advantage: They avoid saddle points. This is illustrated for the two-dimensional case in Fig. 12.27. Three typical paths are shown. Path ① follows the valley to the saddle point, and then proceeds to move to either Ⓐ or Ⓑ. Paths ② and ③ proceed directly to Ⓐ or Ⓑ.

Figure 12.27 points up one of the major difficulties associated with the determination of absolute maximum or minimum values within a given region. If more than one relative maximum or minimum exists, a seek routine will "lock in" on one location, and not find any of the others. An obvious approach toward a solution to this troublesome situation is to select a variety of widely scattered starting points. If enough paths are examined, and if all result in the same location for the optimum, it may be reasonably concluded that the solution found is the absolute optimum within the region in question.

It is to be recognized that the procedures defined in this section can be used also to experimentally solve "optimum problems," even when the functions (or constraints) in question are undefined. At each trial location a sufficient number of closely spaced tests would be carried out. Normally, a minimum of $n + 1$ tests is required, when n independent variables are present. From these, a direction of steepest descent can be ascertained. This is shown diagrammatically in Fig. 12.28 for the case of two independent variables.

Finally, it must be understood that the methods hereinbefore developed presume a euclidean type of geometry. That is, for a given problem the

Figure 12.27



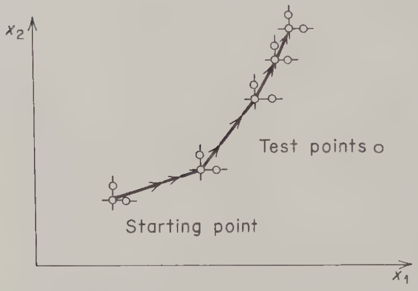


Figure 12.28

coordinate functions—the independent variables—must be of the same dimensions. "Optimum problems" in engineering and applied science, however, frequently contain mixed coordinate variables: length, force, temperature, pressure, cost, etc. To be able to use these methods for each situation, then, it is necessary to nondimensionalize all coordinate quantities. That is not to say that all are nondimensionalized by the same quantities, but that each is related to its appropriate scale factor.

Prob. 12.1: Determine the minimum value of the function

$$f = x_1^2 + 2x_1 + 3x_2^2 + 4(1 - a)$$

Prob. 12.2: In the range $0 \leq x_1 \leq 1.0$ and $0 \leq x_2 \leq 1.0$, determine the maximum and minimum values of the function

$$y = 2x_1^2 + x_1x_2 + 3x_2^2$$

subject to the constraining equation

$$x_1 - x_1x_2 = 0$$

Prob. 12.3: In the range $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 2$, $0 \leq x_3 \leq 3$, $0 \leq x_4 \leq 4$, determine the minimum value of the function

$$y = 3x_1 + 2x_2 + x_3 + x_4^2$$

subject to the following constraining equations:

$$x_1 + 2x_2 = 1 \qquad x_1 + x_3 = 0$$

Prob. 12.4: Find the minimum distance between the two curves

$$y = x^2 - 2x + 2 \qquad \text{and} \qquad y = 1 - x$$

Prob. 12.5: For the three-bar pin-connected indeterminate truss shown in Fig. 12.29, the horizontal component of the deflection at the load point is given by the equation

$$\delta = \frac{QL}{E} \frac{A_1 + 2\sqrt{2}A_2 + A_3}{A_1A_2 + \sqrt{2}A_1A_3 + A_2A_3}$$

The cross-sectional areas of the various members are denoted as A_1, A_2, and A_3. The same material is used throughout the structure, and has a Young's modulus of elasticity of E.

It is desired to obtain for a fixed value of δ the relationships between the cross-sectional areas which correspond to the least total weight of the members of the truss.

In more formal terms, the problem is to determine the minimum value of the linear function φ, subjected to the given nonlinear constraint δ.

$$\varphi = \sqrt{2}x_1 + x_2 + \sqrt{2}x_3$$

$$\delta = \frac{QL^2}{E} \frac{x_1 + 2\sqrt{2}x_2 + x_3}{x_1x_2 + \sqrt{2}x_1x_3 + x_2x_3}$$

where $x_1 = A_1L$
$x_2 = A_2L$
$x_3 = A_3L$

Prob. 12.6: In a two-dimensional space, prove that a straight line is the shortest distance between two points.

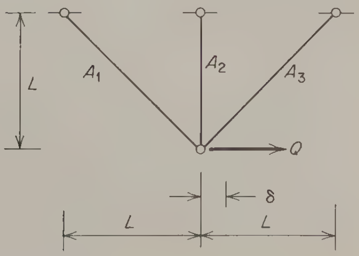
Prob. 12.7: Given the two points $A(1,0,0)$ and $B(0,1,1)$ on the cylindrical surface whose equation is

$$x^2 + y^2 = 1$$

determine the equation of the curve which corresponds to the shortest distance between these two points (see Fig. 12.30).

Prob. 12.8: A column of uniform cross section and length L, pin-connected at one of its ends and rotationally restrained at its other end,

Figure 12.29



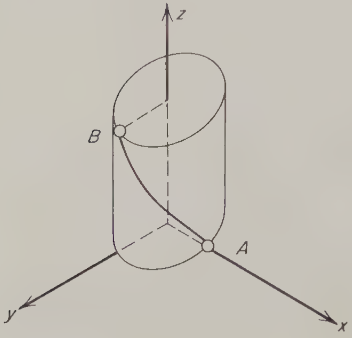


Figure 12.30

is subjected to axial thrusts as shown in Fig. 12.31. The potential energy of the system is given by the equation

$$U = \int_0^L \left(\frac{EI}{2} y''^2 - \frac{P}{2} y'^2 \right) dx + \frac{\beta}{2} y_0'^2$$

where β = restraining spring constant at $x = 0$

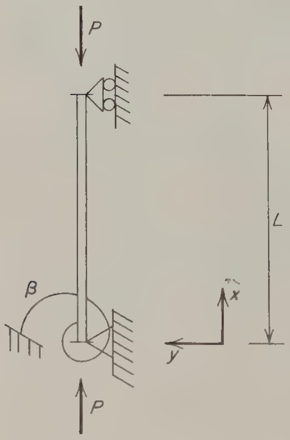
y_0' = slope of member at $x = 0$

The boundary conditions are

$$\begin{aligned} y(0) &= 0 \qquad & EI[y''(0)] + \beta y'(0) &= 0 \\ y(L) &= 0 \qquad & y''(L) &= 0 \end{aligned}$$

Since equilibrium requires that the potential energy of the system be a minimum, determine the value of the critical axial thrust corresponding

Figure 12.31



to the first possibility of lateral deformation, i.e., corresponding to a minimum value of U, by each of the following methods:

(*a*) Using the appropriate Euler equation, and solving the resulting differential equation

(*b*) Using Ritz's method, assuming the smallest-order polynomial that the method and boundary conditions will allow

(*c*) Using Galerkin's method, assuming the smallest-order polynomial that the method and boundary conditions will allow

Prob. 12.9: In the range $1 \leq x_1 \leq 3$, $1 \leq x_2 \leq 3$, $0 \leq x_3 \leq 10$, a function y depends on three independent variables x_1, x_2, and x_3. The following table lists measured values of y at nine different locations.

Point	x_1	x_2	x_3	y
1	+1	+1	+9	+4.0
2	+2	+1	+8	+3.0
3	+3	+1	+7	+2.5
4	+1	+2	+8	+5.0
5	+2	+2	+6	+2.3
6	+3	+2	+4	+3.0
7	+1	+3	+7	+8.0
8	+2	+3	+4	+4.0
9	+3	+3	+1	+5.0

Starting from location $x_1 = 2$, $x_2 = 1$, and $x_3 = 8$ (that is, point 2), using the method of steepest descent, determine the location and value of the minimum y.

index

index

Absolute error, 99
Absolute maxima, 429
Accuracy, analog computer, 7
Adams-Bashford method for analytical continuation, 283
Addition of matrices, 45, 63
Algebraic equations, 66–117, 158–187
  linear, 3
  nonlinear, 3
Analog computer, 5–7
Analytical continuation, 251, 280–286
  ordinary differential equations: collocation method, 285
    predictor-corrector methods, 281–285
Arc length, minimum distance, 447
Array:
  multidimensional, 20
  one-dimensional, 14
  two-dimensional, 14
Augmented groups:
  equations, 437, 444
  matrices, 77

Back substitution, 72
Backward differences, 199, 228
Banachiewicz method (square-root method), 81
Beam-column bending problem, 52, 440
Bessel's integration formula, 232
Bessel's interpolation formula, 210
Biharmonic function:
  equation, 320, 348
  operator, 320
Boundary conditions:
  closed (boundary-value problem), 2
  geometrical, 450
  irregular, 335
  open (initial-and-boundary-value problem), 2
Boundary-value problems, 246, 287–296
  elliptic partial differential equations, 343–379
  nonlinear, 301–314
  relaxation methods (elliptic partial differential equations), 359–379
Branching, 16
Buckling:
  flagpole column, 142
  plate, 352

Calculus of variation, 438–455
Central differences, 227
Characteristic direction:
  hyperbolic differential equations, 403
  nonlinear equations, 416
Characteristic polynomial equation, 138
Characteristic-value problems, 296–301 (*See also* Eigenvalues)
Cholesky's method:
  flow diagram, 115
  solution of inverse problem, 123–127
  solution of linear algebraic equations, 76–81
Closed boundary, 319
Closed integration formulas, 281
Closed region, 319
Cofactors, 25
  reduced, 120
Collatz, L., 455
Collocation method, 256, 285
  analytical continuation, 285
Column buckling, 142
Column vector, 41
Compilers, 8
Computational molecules, 323–335
  laplacian: regular grid, 319, 349, 367
    unequal grids, 376
  polar coordinates, 335
  skewed grid, 333
  square grid differential operators, 327
    higher-order, 328
    volume, 329
  triangular-coordinates, 333, 336
Computers:
  analog, 5–7
    direct, 7
    indirect (mathematical), 7
  digital, 5
  hybrid, 5
  programming, 7–22
    analog, 7
    compilers, 8
    conversational, 8
    destructive read-in, 19
    digital, 5
    Fortran, Mad, and Algol, 8, 19
Computers:
  programming: machine-language coding, 8
    problem-oriented programming languages, 8
Constrained maximum (minimum), 431
Constraining equations, 434
Contour maps, 161, 429
Control theory, 432
Control variables, 427
Convergence:
  iteration for roots of nonlinear equations, 174, 182–184
  parabolic partial differential equations, 385
Convexity, 429
Coordinate transformation, 42, 59
Cramer's rule:
  eigenvalue problems, 137
  solution of inverse problem, 119–122
  solution of linear algebraic equations, 69–71
Crank-Nicolson formulation:
  hyperbolic partial differential equations, 421
  parabolic partial differential equations, 397, 400
Crout's method (*see* Cholesky's method)
Cumulative multiplication, 172
Curve fitting using least-squares method, 233–238

d'Alembert forces, 54, 141
Dc network, 55
Del, 319
Descartes's rule of signs, 164
Destructive read-in, 19
Determinants, 24–39
  cofactor, 25
  methods of evaluation: expansion by minors, 25
    pivotal condensation, 29
    row or column addition, 28
  minor, 25
  pivot, 28
  transpose, 27

Diagonal matrices, 43
Dichotomous search, 460
Difference operators:
  backward, 199
  central, 201
  Δ, ∇, and δ, 197–204
  forward, 198
  interrelationships, table, 202
  shift operator, 198
Difference tables, 188–197, 241–243
  divided, 194
  errors, 195–197
Differences, 190, 222
  backward, 199, 228
  central, 227
  forward, 198, 226, 241
  higher-order, 203
  "off-center" forward, 289
Differential equations, 245–426
  eulerian, 440
    table, 443
  least-squares approximation, 257
Differentiation, 220–229
  tables of equations, 226–228
Digital computer, 5–23
  fixed-point numbers, 14
  floating-point numbers, 14
  flow diagrams, 9
Divided-difference table, 194
Domain, 319

Eigenvalues, 136–157
  flow diagram, computer solution, 155
  ordinary differential equations, 296–301
  partial differential equations, 352, 355
  solution by iteration, 145–147
    largest value, 147
    smallest value, 147
Eigenvector, 138
Elementary row operators, 122
Elimination methods, 71–83
  back substitution, 72
  Cholesky's method, 76–81
  Crout's method, 76
  Gauss' method, 71–76
  inverse problem, 123–127
Elliptic partial differential equations, 319, 343–379
Equations:
  augmented, 437
  characteristic polynomial, 138, 139
  constraining, 434
  contour maps, for nonlinear equations, 161, 429
  differential, 245–426
  homogeneous, 136, 246
  Lagrangian form, 233
  linear, 66, 245
  nonlinear, 66, 158–187, 245
  secular, 138
  transcendental, 175
Errors, 3
  absolute, 3
  approximate methods of solution, 3
  difference tables, 195–197
  double or multiprecision solutions, 3
  inaccuracies in initial data, 3
  interpolation formulas, 217
  lack of precision, roundoff, 3
  propagation, 196
  relative error, 3
  rounding off, 217
  truncation error, 3, 217, 221, 287
Euclid, 460
Euler-Cauchy method, 263
  table, 274
Eulerian differential equation, 440
  table, 443
Euler's method, 262–267
  table, 274
Everett's integration formula, 232
Everett's interpolation formula, 209
Existence of a solution, 2
Explicit recurrence:
  hyperbolic partial differential equations, 418–420
  parabolic partial differential equations, 380–395
Extremum, 438, 451

False-position method, 168–170
Fibonacci numbers, 464, 465
Fibonacci search, 455, 462–465

Finite-difference method, 418
  (*See also* Collocation method)
Fixed-point numbers, 14
Floating-point numbers, 14
Flow diagrams, 9
  branching, 16
  Cholesky's method for obtaining the inverse, 127
  Cholesky's method for obtaining solution of simultaneous linear algebraic equations, 115
  compute box, 9
  counter, 12
  eigenvalue determination, 155
  Gauss-Seidel, solution of simultaneous linear algebraic equations, 103
  looping, 13
  matrix addition, 63
  matrix multiplication, 65
  nesting, 14
  pivotal condensation, evaluation of a determinant, 38, 39
  question box, 10
Forward differences, 198, 226, 241
Functional, 438

Galerkin's methods:
  optimization, 452–455
  solution of ordinary differential equations, 258
Gauss elimination method:
  solution of inverse problem, 126
  solution of linear algebraic equations, 69–76
Gauss-Seidel method for iteration, 85
  flow diagram, 103
Gauss' backward interpolation formula, 209
Gauss' forward interpolation formula, 209
Geometrical boundary conditions, 450
Golden section, 455, 460–462
Gradients, method of, 467–478
Grids, 320–323
  change in size, 371–377
Group relaxation, 92, 364–370
Guest, J., 182

Heun's method, 273
  table, 274
Homogeneous equations, 136, 246
Hybrid computer, 5
Hydraulic network, 57, 367, 377
Hyperbolic partial differential equations, 319, 401–421
  finite difference methods, 418–421
  method of characteristics, 401–418

Identity matrix, 44
Ill-conditioned systems, 68
Implicit recurrence:
  Crank-Nicolson formulation, 397, 400, 421
  hyperbolic partial differential equations, 421
  parabolic partial differential equations, 395–400
Inaccuracies, 3
  (*See also* Errors)
Influencing interval, 194
Initial-value problems, 246
  hyperbolic partial differential equations, 319, 401–421
    explicit recurrence, 418
    implicit recurrence, 421
  ordinary differential equations:
      Euler-Cauchy method, 263
      table, 274
    Euler's method, 262–267
      table, 274
    higher-order equations, 259–261
    Huen's method, 273
      table, 274
    Kutta-Simpson (one-third rule) method, table, 274
    Kutta-Simpson (three-eighths rule) method, table, 274
    Picard's method, 253
    Runge-Kutta methods, 268–280, 287
    Taylor series solution, 247–252, 259, 268
  parabolic partial differential equations, 319, 379–401
    explicit recurrence, 380
    implicit recurrence, 395
  stability, 286, 381, 385

Instability, 286, 381, 385
Integration, 220, 229–233
  Bessel's formula, 232
  closed, 281
  Everett's formula, 232
  formulas, 234–236
  Lagrange formula, 229
  Newton's forward difference formula, 232
  Simpson's one-third rule, 231, 244
  tables of formulas, 234–236
  trapezoidal rule, 230
Interpolation, 204–220
  Bessel's formula, 210
  Everett's formula, 209
  Gauss' backward formula, 209
  Gauss' forward formula, 209
  inverse, 218
  Lagrange formula, 214, 215, 229, 243
  linear, 204
  Newton's backward difference, 207, 209
  Newton's forward difference, 206, 209
  table of formulas, 209–210
Interval of influence, 194
Inverse interpolation, 218
Inverse problem, 118–135
  Cramer's rule, 119–122
  elementary row operators, 122
  elimination methods, 123–127
    Cholesky's method, 127
    Gauss' elimination, 126
  partitioning, 128–131
Irregular boundaries, 335–343
Iteration methods, 83–87, 145, 253–255
  convergence, 174, 385
  "educated guess," 83
  eigenvalues, 145–147
  false-position, 168
  Gauss-Seidel method, 85
  Jacobi's method, 84
  nondiagonal sets, 87
  nonlinear equations, 174, 182–184

Jacobian matrix, 309
Jacobi's method for iteration, 84

Kenneth, P., 304, 307
Kirchhoff's law, 55
Kutta methods, 268–280, 287
Kutta-Simpson method:
  one-third rule, 274
  three-eighths rule, 274

Lagrange interpolation formulas, 214, 215, 229, 243
  errors in, 217
  inverse interpolation, 218
  numerical integration, 229
Lagrangian form of equations, 233
Lagrangian multiplier, 432–438
Language:
  conversational, 8
  Fortran, Mad, and Algol, 8, 19
  machine coding, 8
  problem-oriented, 8
Laplace operator, 319
  change in grid size, 371
  computational molecule, 319, 349, 367, 376
Lattice search, 455
Least-squares approximation:
  curve fitting, 233–238
  solution of differential equations, 257
Leitmann, G., 431
Linear algebraic equations, 66–117
  solution of: Cholesky's method, 76–81
    Cramer's rule, 69–71
    Gauss elimination, 71–76
    Gauss-Seidel iteration, 85
    Jacobi's method, 84
    relaxation methods, 87–94
    square root method, 81–83
Lines, method of, 400
Local maxima, 429
Looping in flow diagramming, 13
Lower triangular matrix, 44

McGill, R., 304, 307
Matrix (matrices), 40–65
  augmented, 77
  column vector, 41

Matrix (matrices):
  diagonal, 43
  general, 4, 40–65
  Jacobian, 309
  lower triangular, 44
  multiplication of: by a constant, 49
    by the identity matrix, 48
  null (zero), 44
  partitioning, 49–51
    bordering, 50
    naturally partitional, 50
    submatrices, 51
  row vector, 41
  square, 20, 43
  transpose, 21, 45
  unit (identity), 44
  upper triangular, 44
  upper unit triangular, 44
Matrix addition, 45, 63
Matrix multiplication, 46, 47, 65
Maxima, 427
  absolute, 429
  constrained, 431
  local, 429
Membrane, natural frequency of vibration, 355
Milne, W. E., 282
Minima, 427
  absolute, 429
  constrained, 431
  local, 429
Minimum distance, 436
Minor, 25
Molecules (*see* Computational molecules)
Multiplication of matrices, 46, 47, 65
Multiplier, Lagrangian, 432, 433

Natural frequency of vibration:
  membrane, 355
  two-mass, spring-connected system, 140
Nesting in flow diagramming, 14
Network: dc, 55
  hydraulic, 57, 367, 377
Newton-Raphson method, 304, 308
Newton's methods:
  backward difference interpolation, 207
  errors in interpolation formulas, 217
  forward difference integration formulas, 232
  forward difference interpolation, 206
  roots of nonlinear equations: first-order method, 170, 186
    second-order method, 171, 181
Nonlinear equations, 66, 158–187
Null matrix, 44
Number sorting, 17

Objective function, 436, 473
"Off-center" differences, 289
Open domain, 319
Open integration formulas, 281
Operator:
  biharmonic, 320
  elementary row, 122
  square grid, differentials, 327–329
Optimization, 427–481
  Galerkin's method, 452
Ordinary differential equations, 245–317
  boundary-value problems, 246, 287, 301–314
  eigenvalues, 296–301
  initial-value problems, 246
Orthogonality method, 258

Paired-trial search, 460
Parabolic partial differential equations, 319, 379–401
Partial differential equations, 318–426
  eigenvalues, 352, 355
  elliptic equations, 319, 343
  hyperbolic equations, 319, 401–421
  parabolic equations, 319, 379–401
Partitioning, 49
  inverse problem, 128–131
Picard's method, 253
Pivotal condensation, 29
  flow diagram, 38
Plate buckling, 352

Poisson's equation, 319, 346
Polar coordinates, 335
  elliptic problem, 349
Polynomial approximations, 255–259
Precision, 3
Predictor-corrector methods, 281–285
Problem types:
  boundary-value, 2, 246, 287
  characteristic-value, 1, 296
  eigenvalue, 1
  initial-value, 2, 246
  inverse, 123
  steady-state, 1
  transient, 1
Propagation of errors, 196
Propagation problems, 319
Proper values, 138
Proportional parts, 204

Quadrature formulas, 231
  (*See also* Integration, formulas)
Quasi-linear partial differential equations, 319

Raphson-Newton method, 304, 308
Recurrence:
  explicit, 380, 418
  implicit, 395, 397, 400, 421
Reduced cofactor, inverse problem, 120
Region:
  closed, 319
  open, 319
Relative error, 3
Relaxation methods, 87–94
  group-, 92, 364–370
  residuals, 89
  unit-, 89, 360–364
Residuals, 89, 361
Ritz's method, 449–452
Roots of equations:
  approximations: largest, 163
    smallest, 164
  conjugate complex, 162, 177
  nonlinear equations, 170, 171, 181, 186
  real, 161, 180
Rope, deflection of, 446
Round off, 3
  errors, 217
Round off error (*see* Errors)
Row operators, elementary, 122
Row vector, 41
Runge-Kutta methods:
  method of lines, 400
  ordinary differential equations, 268–280, 287

Saddle point, 430
Scale factors, 93, 478
Scanning (nonlinear equations), 159
Search methods, 455–478
  multidimensional problems, 466–478
    method of gradients, 467, 477
    one-at-a-time method, 456, 466
    pattern search, 456
    sectioning, 456, 466
    steepest ascent, 455, 467
  single-variable problems, 455–465
    dichotomous search (paired-trials), 460
    Fibonacci search, 455, 462–465
    Golden section, 455, 460–462
    lattice search, 455
Search plan, 457–466
Secular equation, 138
Seek methods, 455
  (*See also* Search methods)
Series, transcendental functions, 175
Sets of equations:
  algebraic, 66
  first-order ordinary differential equations, 252, 254, 275
  higher-order ordinary differential equations, 259, 278
  nonlinear, 180
Shift operator, 198
Simpson's one-third rule for integration, 231, 244
Simpson's three-eighths rule for integration, 274
Spring-connected system:
  natural frequency of vibration, 140
  two-mass, 53, 54

Spurious solutions, 287
Square root (Banachiewicz) method, 81–83
Square roots, 23
Stability, numerical, initial-value problems, 286, 381, 385
Star points, unequal leg, 337–343, 371
Starting series, 251
Stationary point, 430, 442
Steady-state problems, 1
Steepest descent:
  method of, 467–478
  table of equations, 474
Stencils, 324
  (*See also* Computational molecules)
Subdivision method, 257
Subscripted numbers:
  multidimensional arrays, 20
  one-dimensional array, 14
  two-dimensional arrays, 14
Subsidiary equations, 434
  (*See also* Constraining equations)
Summation:
  double, 11
  products, 11
Synthetic division, 166–168
Systems:
  continuous, 2
  deterministic, 433
  discrete element, 1
  ill-conditioned, 68
  lumped, 2
  nonlinear, 416
  static, 433
  stochastic, 433

Tangents, method of, 170, 171
Taylor series solutions, ordinary differential equations, 247–252, 259, 268
Transcendental equations, real roots, 175
Transformation, coordinate, 42, 59
Transpose:
  determinant, 27
  matrices, 45
Trapezoidal integration rule, 230
Truncation errors, 3, 217, 221, 287
Truss, pin-connected: three-bar plane, 473
  three-bar space, 56
  two-bar plane, 434

Unimodality, 429, 456
Unit matrix, 44
Unit relaxation tables, 360
Upper triangular matrix, 44
Upper unit triangular matrix, 44

Variation, calculus of, 438–455
Vector:
  column, 41
  row, 41

Wave propagation, 413

Zero matrices, 44

This book was set in Modern by The Maple Press Company, and printed on permanent paper and bound by The Maple Press Company. The designer was Merrill Haber; the drawings were done by J. & R. Technical Services, Inc. The editors were B. J. Clark and Dorothy Graham. William P. Weiss supervised the production.

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

KEN NAGELE
6390 DRY FORK RD.
CLEVES, OHIO 45002

34423